Session 3 - Comparison of Populations

Brian Leroux

Wednesday, January 23, 2019

Outline

1. Analysis of Experiments for Comparison of Two Population Means

- Randomized experiments versus observational studies
- The Z-test for large samples
- Confidence intervals for a difference between means

2. The 2-sample t-test

- the assumption of equal variances
- assumption of normality of the populations

3. Analysis of Experiments for Comparison of Two Proportions

- Z-test and chi-squared test
- Confidence intervals for a difference between proportions
- Continuity correction
- Fisher's exact test

1. Analysis of Experiments for Comparison of Two Population Means

- Randomized experiments versus observational studies
- The Z-test for large samples
- Confidence intervals for a difference between means

Example: an agricultural experiment

An agricultural researcher wants to compare the expected yields from two different fertilizers (A and B). An experiment is performed using 20 1-acre plots of land. A random sample of 10 of the plots is chosen and fertilizer A is applied to these, with fertilizer B applied to the remaining 10 plots. The crop yield (in tons) is measured at the end of the season for each plot.

We define the null hypothesis as

$$H_0 : \mu_A = \mu_B,$$

where μ_A and μ_B are the mean yields per plot for fertilizers A and B, respectively. If we have no reason to believe one of the fertilizers is more likely to be the superior one, we set the alternative hypothesis to be

$$H_1 : \mu_A \neq \mu_B.$$

Note: we think of μ_A as the mean of the *population* of yields for fertilizer A from a hypothetical infinite population of 1-acre plots of land.

Type I and Type II Errors

Just as in the 1-sample hypothesis testing situation seen last week, there are two types of errors that can be made:

- Type I error: reject H_0 when it is true
- Type II error: fail to reject H_0 when it is false (recall that we don't say "accept H_0")

We typically want to control the Type I error probability (e.g., 0.05) while also minimizing the Type II error probability (maximizing the power) as far as possible.

Questions about the experimental design

1. Is 20 plots enough?

- Consider power calculations

2. What if fertilizer A happens to get assigned to the 10 most fertile plots?

- Consider alternative designs (e.g., randomized block design, in which randomization is done separately within high fertility and low fertility plots)

Example: A/B Testing

A company wants to compare two design versions for a web page to determine which one yields higher sales. During a given period the first 2000 customers who visit the page are randomly assigned to see either version A or version B. The total amount of sales resulting from each visit is recorded.

As with the agricultural experiment, we want to test $H_0 : \mu_A = \mu_B$ vs. $H_1 : \mu_A \neq \mu_B$, where μ_A and μ_B are the mean sales per visit for the two versions. Again we would like to control type I error probability and maintain adequate power.

Questions about the A/B Testing Design

1. Why 2000 customers?

- Does it provide adequate power?
- If it is more than needed for adequate power, consider using a smaller number so that more experiments can be performed, e.g., to test different versions of the web page.

2. Is the answer the same for different types of customers (e.g., by location)?

- Consider randomized block design in which randomization is done separately by location.

3. What about combinations of different web page features?

- Consider factorial designs in which all combinations of features are used. For example, if the features are background color (red versus blue) and font size (large versus small), a factorial design would include all 4 combinations (red/large, red/small, blue/large, blue/small).

Randomized Experiments

The most important feature of experiments such as the examples given above is the *random* assignment of conditions to the experimental units.

- Random assignment of fertilizers to plots of land
- Random assignment of web page versions to users

Randomization is the basis for statistical inference - it allows us to make inference in terms of probability statements about the population means.

Randomization also ensures an unbiased comparison - on average the results will reflect the true population means.

Randomized Versus Observational Studies

Suppose that in A/B testing, a web-site visitor is allowed to select one or the other version of the page. We could not use the results to infer that one version is better than the other. The types of people who choose version A might be the types of people who spend more than those who choose version B.

This is an example of *bias* in observational studies.

A randomized design is preferable to an observational study whenever randomization is feasible.

When randomization is not feasible

Suppose we want to compare sales from men and women visiting a web site. We can't randomize their gender! Or suppose we want to compare sales from visitors on Saturday compared with Sunday. We don't control who visits on the different days so we must use an observational study.

However, we can still set up the hypothesis in the same way. For the Saturday/Sunday comparions, we would test $H_0 : \mu_{\text{Sat}} = \mu_{\text{Sun}}$.

In this case the means to be compared are the means of the populations of all visitors on the given day of the week.

For randomized experiments and observational studies, the *goal* is the same – compare 2 population means. However, the *interpretation* of the results is different because of the potential bias in an observational study. In addition, different *statistical analysis methods* may be used to try to address the bias (e.g., stratification, regression).

Hypothesis Testing for Comparing Two Means

Test $H_0 : \mu_A = \mu_B$ versus $H_1 : \mu_A \neq \mu_B$.

The Data: 2 independent samples of observations, one sample from each population

1. $X_{A1}, X_{A2}, \ldots, X_{An_A}$ from population A with mean μ_A
2. $X_{B1}, X_{B2}, \ldots, X_{Bn_B}$ from population B with mean μ_B

Note: the sample sizes n_A and n_B don't have to be equal.

Test Statistic

Because we wish to compare population means, our hypothesis test is based on the difference between the two sample means $\bar{X}_A - \bar{X}_B$, where

$$\bar{X}_A = \frac{1}{n_A} \sum_{i=1}^{n_A} X_{Ai}$$

$$\bar{X}_B = \frac{1}{n_B} \sum_{i=1}^{n_B} X_{Bi}$$

How do we assess the statistical significance of the difference? - Use the *sampling distribution* of the difference between sample means.

The Sampling Distribution of the Difference Between Two Sample Means

What do we know from the 1-sample situation: each sample mean has expected value equal to its respective population mean and variance equal to its respective population variance divided by the sample size, i.e.,

$$E(\bar{X}_A) = \mu_A,$$

and

$$\text{var}(\bar{X}_A) = \sigma_A^2/n_A$$

and similarly for $\bar{X}_B$.

By a property of expected values of random variables, the expected value of the difference between the sample means is equal to the difference between their means:

$$E(\bar{X}_A - \bar{X}_B) = \mu_A - \mu_B.$$

So it is an unbiased estimator of the difference between population means.

The variance of the difference between two independent sample means

The variance of the *difference* between the sample means is equal to the *sum* of their variances:

$$\text{var}(\bar{X}_A - \bar{X}_B) = \text{var}(\bar{X}_A) + (\bar{X}_B)$$

Therefore:

$$\text{var}(\bar{X}_A - \bar{X}_B) = \sigma_A^2/n_A + \sigma_B^2/n_B.$$

Note: This requires that the samples are *independent* (more on this later).

The variance of a DIFFERENCE is equal to the SUM of the variances

Demonstration by simulation (results are approximate due to random simulation error)

$\text{var}(X - Y) = \text{var}(X) + \text{var}(Y)$ if X and Y are independent

```
set.seed(2)
x=rnorm(10,0,1)
y=rnorm(10,0,1)
d=data.frame(x,y,d=x-y)
round(d,2)
```

```
       x     y     d
1  -0.90  0.42 -1.31
2   0.18  0.98 -0.80
3   1.59 -0.39  1.98
4  -1.13 -1.04 -0.09
5  -0.08  1.78 -1.86
6   0.13 -2.31  2.44
7   0.71  0.88 -0.17
8  -0.24  0.04 -0.28
9   1.98  1.01  0.97
10 -0.14  0.43 -0.57
```

```
apply(d,2,var)
```

```
         x          y          d
0.9702007 1.3948335 1.8998614
```

Constructing the Test Statistic

We want a test statistic that has an approximate standard normal distribution (or maybe a t distribution if the sample sizes are small). In the 1-sample case we used $Z = (\bar{X} - \mu_0)/(s/\sqrt{n})$, or $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ if we know σ. We have the following general recipe for a test statistic that works in a wide variety of settings (e.g., 2-sample comparisons, linear regression, etc.):

$$\text{Test Statistic} = \frac{\text{Sample Estimate} - \text{Hypothesized Value}}{\text{SE of the Sample Estimate}}$$

The sample estimate is $\bar{X}_A - \bar{X}_B$ and the hypothesized value is 0.

The Two-Sample Z-Statistic

Assuming the variances are unknown (as is usually the case), we substitute their respective sample variances in the standard error formula to get the following test statistic:[1]

$$Z = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{s_A^2/n_A + s_B^2/n_B}}.$$

To determine statistical significance, we use similar rules as in the 1-sample case.

[1] If we do happen to know the population variances we would use them in the standard error formula but this is very rare.

Assessing Statistical Significance

There are 3 cases to consider.

1. Large sample sizes (both n_A and n_B): in this case, we compare Z to the standard normal distribution. ("the large-sample Z-test")
2. Small sample sizes (either or both) with assumed equal variances: in this case we use the `equal-variance t-test`.
3. Small sample sizes (either or both) without assuming equal variances: in this case use the `Welch t-test`. (Warning: there are other names for this test.)

Independence Assumptions

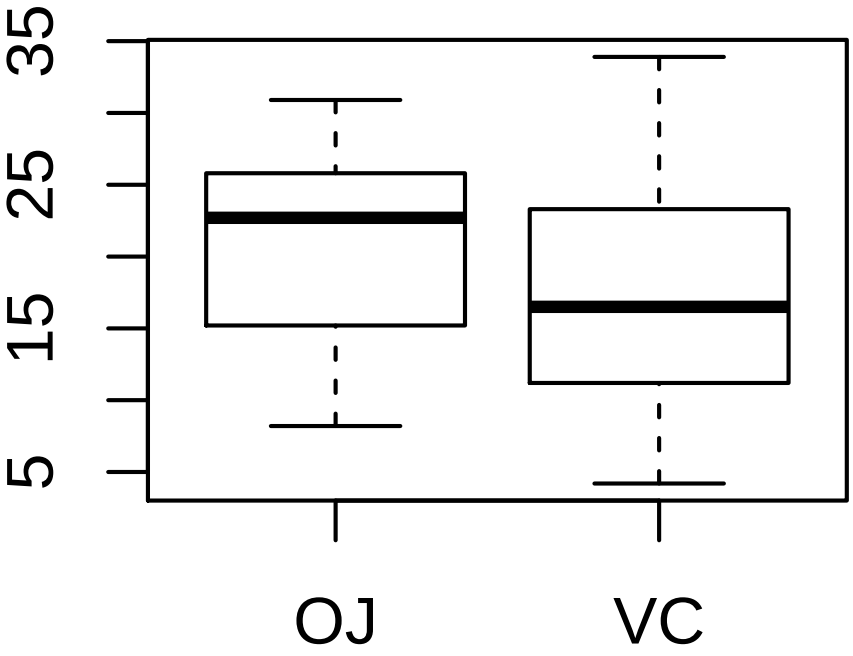
In all three cases, there are two critical independence assumptions:

1. The samples are independent (e.g., you can't use the same experimental units in the two groups).
2. The observations within each sample are independent. This is another way of saying that we can think of each sample as a simple random sample from the corresponding population.

Case 1: Vitamin C and Tooth Growth ('ToothGrowth' data)

The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received vitamin C in one of 2 forms: orange juice (OJ), or ascorbic acid (VC). (There also was a dose factor in the experiment but that will be ignored for now.)

```
boxplot(split(ToothGrowth$len,ToothGrowth$supp))
```



Let's assume for now that the sample sizes are large enough to use the large-sample Z-test.

Comparing mean tooth lengths in the groups

```
m=with(ToothGrowth,tapply(len,supp,mean))
s=with(ToothGrowth,tapply(len,supp,sd))
n=with(ToothGrowth,tapply(len,supp,length))
data.frame(m,s,n)
```

```
          m        s  n
OJ 20.66333 6.605561 30
VC 16.96333 8.266029 30
```

```
z=(m[1]-m[2])/sqrt(sum(s^2/n))
data.frame(z,p=round(2*(1-pnorm(z)),4))
```

```
          z      p
OJ 1.915268 0.0555
```

We would not reject the null hypothesis of equal means at the 0.05 level of significance (p=0.06).

Confidence interval for the difference between population means

```
se=sqrt(s[1]^2/n[1]+s[2]^2/n[2])
z.05=qnorm(0.975)
lower = m[1]-m[2]-z.05*se
upper = m[1]-m[2]+z.05*se
lower
```

```
         OJ
-0.08634516
```

```
upper
```

```
      OJ
7.486345
```

Next - how to deal with the small sample situation

2. The 2-sample t-test

- the assumption of equal variances
- assumption of normality of the populations

Case 2: The equal-variance t-test

If sample sizes are small one option is to use the equal-variance t-test. This test assumes that the **variances for the two populations are equal** and that **the population distributions are normal**. If these assumptions are true the sample sizes do not have to be large.

The test statistic uses a single "pooled" estimate of the variance which is a weighted average of the two sample variances:

$$s^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A - 1 + n_B - 1}$$

The test statistic Z then becomes

$$Z = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{s^2/n_A + s^2/n_B}} = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{s^2(1/n_A + 1/n_B)}}$$

The p-value is calculated using the $t_{n_A+n_B-2}$ distribution.

Note that when the sample sizes are equal the test statistic will be the same whether the separate variances are used or the pooled variance is used.

The equal-variance test for comparing OJ and VC groups

```
pooled.sample.var=sum((n-1)*s^2)/sum(n-1)
z=(m[1]-m[2])/sqrt(pooled.sample.var*sum(1/n))
data.frame(z, p=round(2*(1-pt(z,df=sum(n)-2)),4))
```

```
          z      p
OJ 1.915268 0.0604
```

The conclusion is the same as for the large-sample test - we do not reject the null hypothesis.

Confidence interval for the difference between population means with equal variances assumed

```
se=sqrt(pooled.sample.var*sum(1/n))
t.05=qt(0.975,df=sum(n)-2)
lower = m[1]-m[2]-t.05*se
upper = m[1]-m[2]+t.05*se
print(lower)
```

```
        OJ
-0.1670064
```

```
print(upper)
```

```
      OJ
7.567006
```

The confidence interval is (-0.17, 7.57), which is very similar to the result from the large-sample procedure.

Peforming the equal-variance t-test using the 't.test' function

Set the `var.equal` option to TRUE:

```
options(width=56)
with(ToothGrowth,
t.test(len[supp=="OJ"],len[supp=="VC"],var.equal=T))
```

```
    Two Sample t-test

data:  len[supp == "OJ"] and len[supp == "VC"]
t = 1.9153, df = 58, p-value = 0.06039
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1670064  7.5670064
sample estimates:
mean of x mean of y
 20.66333  16.96333
```

Case 3: The Welch t-test

This test uses an adjusted degrees of freedom:

```
welch.df = (s[1]^2/n[1] + s[2]^2/n[2])^2/
(s[1]^4/(n[1]^2*(n[1]-1)) + s[2]^4/(n[2]^2*(n[2]-1)))
welch.df
```

```
      OJ
55.30943
```

This is similar to the value used for the equal-variance t-test, which is often the case particularly when group sample sizes are equal.

In general, a "2-sample t-test" can refer either to the equal-variance version or the Welch version. They often agree quite closely.

Warning: if the sample variances and/or sample sizes are very different, the result can be quite sensitive to choice of test. In practice, if the two tests give conflicting conclusions there is probably something else going on that casts doubt on the validity of either test and alternatives should be considered (e.g., permutation test).

95% confidence interval based on the Welch df

```
se=sqrt(sum(s^2/n))
t.05=qt(0.975,welch.df)
lower = m[1]-m[2]-t.05*se
upper = m[1]-m[2]+t.05*se
lower
```

```
        OJ
-0.1710156
```

```
upper
```

```
      OJ
7.571016
```

The result is very similar to the equal-variance confidence interval.

Peforming the Welch t-test using the 't.test' function

Set the `var.equal` option to FALSE:

```
options(width=56)
with(ToothGrowth,
t.test(len[supp=="OJ"],len[supp=="VC"],var.equal=F))
```

```
    Welch Two Sample t-test

data:  len[supp == "OJ"] and len[supp == "VC"]
t = 1.9153, df = 55.309, p-value = 0.06063
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1710156  7.5710156
sample estimates:
mean of x mean of y
 20.66333  16.96333
```

Summary of tests for comparing two population means

We have two *independent* random samples from populations with means μ_A and μ_B and unknown[2] variances σ_A^2 and σ_B^2. We want to test $H_0 : \mu_A = \mu_B$ versus $H_1 : \mu_A \neq \mu_B$.

1. Calculate the test statistic: $Z = (\bar{X}_A - \bar{X}_B)/\sqrt{s_A^2/n_A + s_B^2/n_B}$
2. Calculate the critical value for the desired significance level (α) and/or the p-value using one of three cases:

- 1. If n_A and n_B are *both* large, use the standard normal distribution.
- 2. If n_A or n_B is small *and* the population distributions are both normal, *and* $\sigma_A^2 = \sigma_B^2$, then use the $t_{n_A+n_B-2}$ distribution.
- 3. If n_A or n_B is small *and the population distributions are both normal*, and $\sigma_A^2 = \sigma_B^2$ cannot be assumed, then use the t distribution with the Welch df.

Otherwise, (e.g., with small sample sizes and possibly non-normal distributions), we are stuck for now (but will return to this topic later).

[2] if the variances are known use the true values in place of the sample variances.

Summary of assumptions for the two-sample t-test

There are three requirements (in order of importance):

1. Independence: the two samples must be independent and the observations within each sample must be independent.

If the samples are not independent (e.g., they are paired), or if the observations within each sample are not independent, then none of the tests are valid even if the sample sizes are large. Specialized methods are needed.

2. Equal variances (if using the equal-variance t-test)

If the variances are not equal, the equal-variance t-test is not valid even if the sample sizes are large.

3. Sample sizes must be large *or* the population distributions must be normal.

In the large sample case, we may refer to the test as either a `Z-test` or a `t-test`

The three assumptions on the previous slide are listed roughly in order of importance. Non-independence (either of the samples or of observations within samples) can have a huge effect on the test performance *even if the sample sizes are large*. The effect can be to make the test either overly conservative (type I error probability too small) or anti-conservative (type I error probability too large).

Non-equal variances invalidate the equal-variance t-test (it can be either conservative or anti-conservative). This is true *even if the sample sizes are large*; hence it is never advisable to assume equal variances with large samples.

The normality assumption is the least important. The test performs well for any distribution if the sample sizes are large enough, and the test is quite robust to departures from normality for moderate sample sizes.

How do i decide which test to use?

The goal is to obtain a valid test, i.e., if i set the significance level at a certain level α then the test will reject the null hypothesis *when it is true*, with probability equal to α.

With large sample sizes, use the large-sample test and do not assume equal variances. Note that with large sample sizes (so large df) the t and normal distributions will be very similar so it doesn't matter whether you use the normal critical values or the t-distribution critical value. The t critical value, t_{α,n_A+n_B-2}, will be very close to the normal critical value z_α.

With small sample sizes, you first have to decide if you can assume that the population distribution is normal. If yes, then you decide whether you can assume the population variances are equal (if yes, use the equal-variance test; if no, use the Welch test).

How do I decide whether to assume equal variances or not?

There might be a theoretical justification for the assumption.

There might be extensive experience with similar experiments that have demonstrated equal variances and the experimental set-up has not changed.

A third possibility is to use the data to help decide. This is problematic because we are in the small-sample situation, which usually means we don't have enough data to properly assess whether the variances are equal or not. (see examples in later slides.)

An ill-advised approach

One approach presented in some (older) statistics texts is to *perform a hypothesis test* (an F-test) to compare the variances and then use the result of that test to guide your decision. This is **not recommended** for the following reasons:

1. When the sample sizes are small (which is when the equal variance assumption is relevant), the test to compare variances has very little power. Therefore, this approach will often lead you to assume equal variances even when that assumption is not reasonable.
2. When the sample sizes are large, the test to compare variances may have adequate power. However, in this case, there is no benefit in assuming equal variances and you might as well always use unequal variances.
3. The F-test for comparing variances is highly sensitive to non-normality. Hence, it can give misleading results if the distribution is not normal. (On the other hand, the t-test is relatively robust to non-normality).

How do I decide whether or not I can assume normal populations?

The answer is similar to that for the equal-variance assumption.

1. A theoretical justification or extensive experience from past experiments is best.
2. We can use the data to assess the assumption, but again, we are in the small sample situation where we likely don't have enough data to do a good job.
3. Performing a test for normality is not advised for similar reasons as for the F-test for comparing variances.

Using the data (e.g., q-q plots) is reasonable to help make a judgement about the normality assumption.

Residual Diagnostics

Methods for using the data to check assumptions of equal variances and normality use the concept of `residuals`. A general definition of a residual (which applies to ANOVA, linear regression and other methods) is the following.

> *A **residual** is the difference between an observation and its predicted value based on a statistical model.*

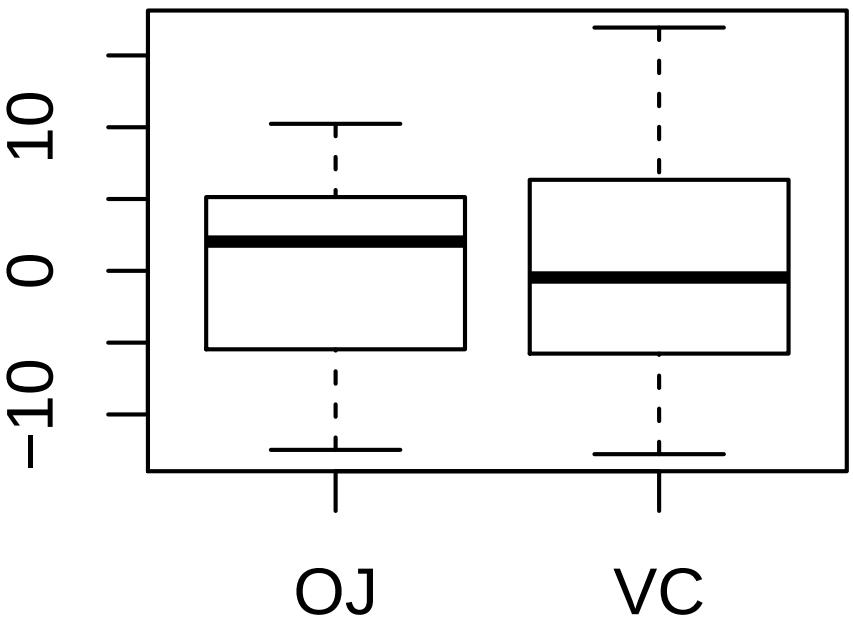
In the 2-group comparison situation, the residual for each data value is simply the data value minus its group sample mean.

Residuals allow us to combine data from the two groups to get more sensitive checks on the model assumptions. For example, we can assess normality overall rather than checking it in each group separately. This becomes increasingly important for ANOVA and regression.

Residuals to Check Equality of Variances

The equality of variances can be assessed graphically using a boxplot (or dotplot).

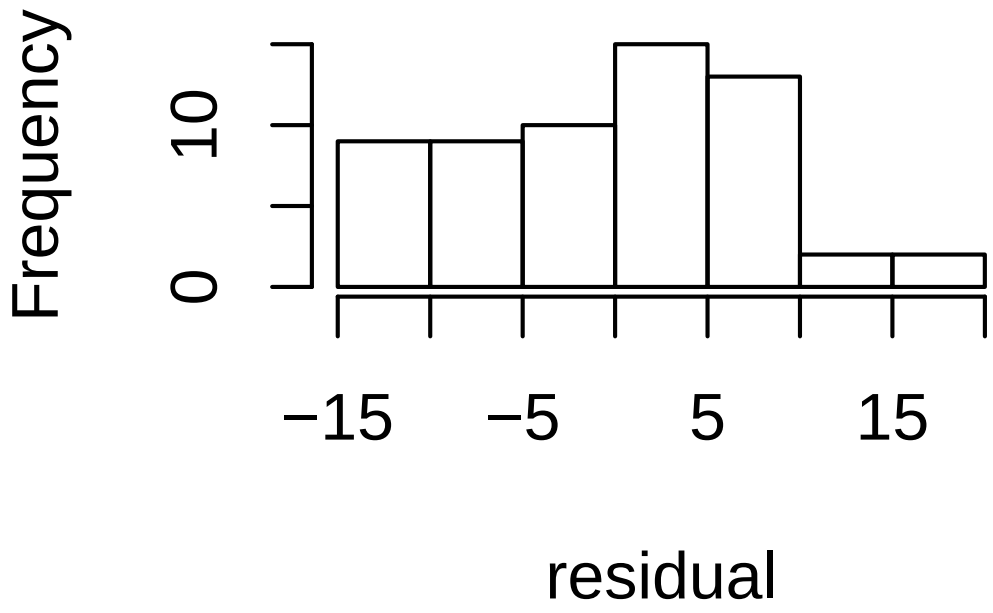
```
attach(ToothGrowth)
group.means = tapply(len,supp,mean)
residual = len - group.means[tapply(len,supp)]
boxplot(split(residual,supp))
```



This is similar to the boxplot of the data except that both groups have been centered to have mean 0. This plot would be interpreted as indicating no evidence against equality of variances.

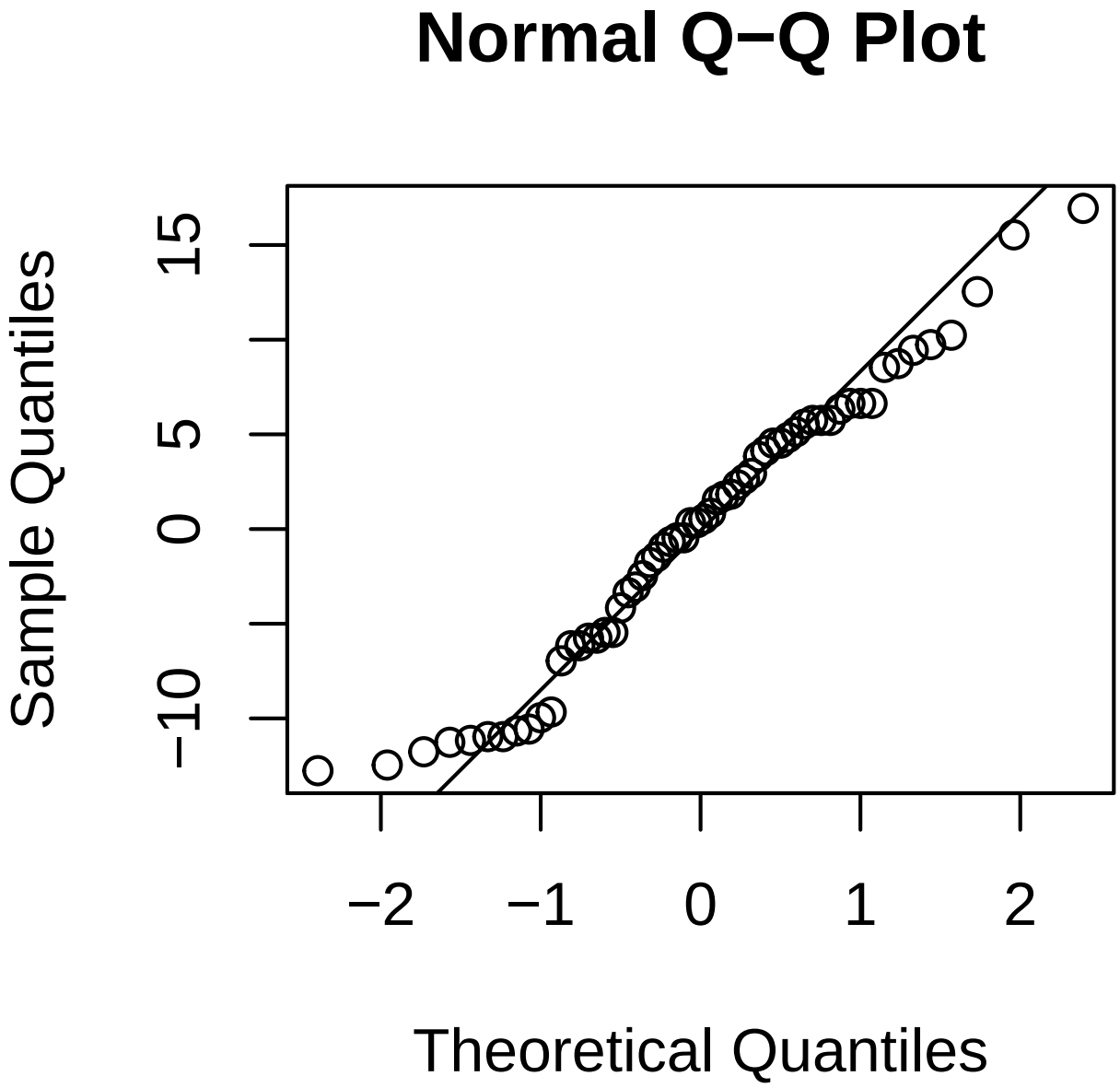
Using Residuals to Check Normality

We can also use the residuals to check normality by applying the usual graphical methods (histograms and q-q plots) to the residuals.



The histogram looks somewhat non-normal, but recall that assessing normality with small data sets is difficult. The ultimate decision is based on whether the magnitude of departure from normality is large enough to invalidate the test – this requires simulation studies on the robustness of the t-test.

The Q-Q Plot of the Residuals



Again, the q-q plot does not clearly indicate non-normality, but a clear conclusion is not possible.

How do I decide if the independence assumptions are valid?

Independence is an assumption that is verified through our knowledge of the study design and data collection mechanisms. We typically do not use the data to check on the assumption of independence.

For example, in a "completely-randomized" experimental design, we randomly assign the experimental units to the different conditions completely randomly. This design is often used in randomized clinical trials, in which each patient is randomly assigned to one of the treatments under study using a random number generator in such a way that their treatment assignment is not influenced by the assignments of any other patients in the trial. For a completely randomized design, the method of randomization justifies the independence assumptions.

Similarly, in an experiment comparing quality of manufactured items using three different processes, we would take random samples of items from each process in such a way that the sampling was done independently for each process.

Indication of Non-Independence: Blocked Designs

Blocking is one of the ways that the assumption of independence can be violated.

In an experimental situation, sometimes we use *blocking* (also called stratifying) to control variability and increase the precision of our experiment. For example, in multi-center clinical trials, each "center" (e.g., a clinic or hospital) is treated as a homogeneous block and we would randomize patients to treatment groups using a separate random number list within each block. This means that the observations are no longer independent, e.g., outcomes for patients in the same center would be expected to be correlated (not independent).

In a quality testing experiment, we might use 10 different batches of raw material and for each batch make one run of each process using that same batch of material. Thus, the batches form the blocks and the outcomes for two units produced using the same batch might be expected to be correlated (not independent).

We will come back to the question of how to analyze data from designs that use blocking. For now, we will assume independence assumptions are satisfied and consider how to assess the other two assumptions of ANOVA.

Next - comparison of population proportions

$$H_0 : p_A = p_B$$

3. Analysis of Experiments for Comparison of Two Proportions

- Z-test and chi-squared test
- Confidence intervals for a difference between proportions
- Continuity correction
- Fisher's exact test

Example: Randomized Clinical Trial

In a clinical trial, we might wish to compare the proportions of patients who survive when given a new treatment compared with an older treatment. We would want to test the null hypothesis that the probability of surviving is the same under the two treatments versus the alternative hypothesis that the survival probabilities are not equal.

Note: even if the new treatment is thought to be better than the old, we typically use a 2-sided alternative hypothesis because we want to be able to detect if the new treatment is in fact inferior (perhaps due to unsuspected side effects).

Randomly assinging patients to one or the other treatment will ensure an unbiased estimate of the difference in survival probabilities and allow us to make a valid test of the hypothesis.

Observational Medical Studies

If it is not feasible to randomly assign treatments to patients, then we could perform an observational study by comparing results for those patients who choose to receive the new treatment versus those who choose the old treatment. This suffers from bias due to the fact that patients who choose one treatment are likely different than patients who choose the other treatment in ways that may affect their outcome (for example, sicker patients might choose the new treatment).

Epidemiological studies of health effects of environmental or dietary exposures are examples of this type of study. For example, we might compare incidence rates of cancer between people with high versus low fat diets. The possibility of bias in such a study is very high. Specialized regression methods have been developed to try to control such bias.

Example: A/B Testing with Dichotomous Response

In A/B testing we might record whether or not a visitor to a web site follows a certain link.

We would like to test the null hypothesis that the proportion of users who select the link is the same in the two versions of the site.

To design the experiment we need a test of the null hypothesis that has a fixed type I error and with adequate power to detect a meaningful alternative hypothesis.

Comparison of Proportions from Two Independent Samples

We need a test procedure that is analogous to the 2-sample t-test to compare two population proportions from independent samples. The special nature of the data (a binary variable that takes values 0 or 1) means that we have to modify our approach for comparing means.

Suppose that we have two independent random samples of binary data with population proportions p_A and p_B. We want to test $H_0 : p_A = p_B$ against the alternative hypothesis $H_1 : p_A \neq p_B$.

We base our test on the difference between the sample proportions $\hat{p}_A - \hat{p}_B$, where the sampling proportions are the proportions of 1s in each sample. Thus, we need to determine the sampling distribution of this statistic.

The Sampling Distribution of the Difference Between Two Sample Proportions

We know from before that each of the sample proportions has expected value equal to its respective population proportion, i.e.,

$$E(\hat{p}_A) = p_A,$$

and variance equal to

$$\text{var}(\hat{p}_A) = p_A(1 - p_A)/n,$$

and similarly for $\hat{p}_B$.

As before, using the properties of expected values and variances of sums of independent random variables, we get

$$E(\hat{p}_A - \hat{p}_B) = p_A - p_B.$$

and

$$\text{var}(\hat{p}_A - \hat{p}_B) = p_A(1 - p_A)/n_A + p_B(1 - p_B)/n_B.$$

Using the Pooled Sample Proportion

So far, things are following the same pattern as for the t-test for comparison of means. However, the estimation of the variance of the difference between sample proportions has a special form compared to what we had for the difference between sample means. Because of the special $p(1-p)$ form of the variance of a binary variable, we don't have any σ^2 parameters to deal with. This means that *under the null hypothesis*, which says that $p_A = p_B$, the variance formula takes the special form

$$\text{var}(\hat{p}_A - \hat{p}_B) = p(1-p)(1/n_A + 1/n_B),$$

where p is the common (under H_0) value of p_A and p_B.

This means that we must estimate an assumed common value of p. This is done simply by calculating the overall proportion of 1s in the pooled sample, i.e.,

$$\hat{p} = \frac{n_A \hat{p}_A + n_B \hat{p}_B}{n_A + n_B}$$

Thus, we have our estimate of the variance of the difference between the sample proportions equal to $\hat{p}(1-\hat{p})(1/n_A + 1/n_B)$, which implies that the estimated SE of the difference between proportions is

$$SE(\hat{p}_A - \hat{p}_B) = \sqrt{\hat{p}(1-\hat{p})(1/n_A + 1/n_B)}$$

The Test Statistic for Comparison of Proportions

The test statistic follows the general recipe we saw before:

$$\text{Test Statistic} = \frac{\text{Sample Estimate} - \text{Hypothesized Value}}{\text{SE of the Sample Estimate}}$$

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})(1/n_A + 1/n_B)}}.$$

The rules for interpreting Z are similar to those for the comparison of sample means, except that there is no consideration of normality for the population.

Interpretation of Z

Summary of Tests for Comparison of Two Population Means from Independent Samples

We have two *independent* random samples from populations with proportions p_A and p_B. We want to test $H_0 : p_A = p_B$ versus $H_1 : p_A \neq p_B$.

1. Calculate the test statistic:

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})(1/n_A + 1/n_B)}}.$$

2. Calculate the critical value for the desired significance level (α) using the standard normal distribution, (e.g., 1.96) for $\alpha = 0.05$.

This test requires large sample sizes. We call it the large-sample Z-test for comparing proportions.

Example: A Randomized Clinical Trial

Suppose that an experiment was done to compare two different types of CPR for victims of cardiac arrest and had the following results: 1000 patients were randomized to treatment A, of whom 100 survived, and 1050 were randomized to treatment B, of whom 90 survived. Is there evidence that the probability of survival is different for the two treatments?

```
n = c(1000,1050)
y = c(100,90)
phat=y/n
diff=diff(phat)
pooled.p=sum(n*phat)/sum(n)
se=sqrt(pooled.p*(1-pooled.p)*sum(1/n))
z=diff/se
data.frame(phat[1],phat[2],diff,se,z,p=2*(1-pnorm(abs(z))))
```

```
  phat.1.    phat.2.         diff          se         z
1     0.1 0.08571429 -0.01428571 0.01281332 -1.114911
          p
1 0.2648885
```

There is not evidence that the two survival probabilities are different.

95% confidence interval for the difference between survival probabilities

```
lower = diff-1.96*se
upper = diff+1.96*se
lower
```

```
[1] -0.03939982
```

```
upper
```

```
[1] 0.01082839
```

Comparing proportions with the R function 'prop.test'

This function is based on the cross-tab of treatment group and survival.

```
cross.tab=data.frame(rbind(y,n-y))
names(cross.tab)=c("A","B")
row.names(cross.tab)=c("survived","died")
cross.tab
```

```
            A   B
survived  100  90
died      900 960
```

```
prop.test(cbind(y,n-y),correct=F)
```

```
    2-sample test for equality of proportions
    without continuity correction

data:  cbind(y, n - y)
X-squared = 1.243, df = 1, p-value = 0.2649
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01086266  0.03943409
sample estimates:
    prop 1     prop 2
0.10000000 0.08571429
```

The Chi-Squared Test

The chi-squared test is a general test procedure for testing hypotheses with categorical data. In the simple case of comparing proportions the chi-squared test is equivalent to the Z test for comparing proportions.

```
chisq.test(cbind(y,n-y),correct=F)
```

```
    Pearson's Chi-squared test

data:  cbind(y, n - y)
X-squared = 1.243, df = 1, p-value = 0.2649
```

Note that the p-values for the two tests are the same. The test statistic "X-squared" is equal to the square of the Z statistic. The chi-squared test is useful for more general types of hypothesis testing, such as comparison of 3 or more population proportions.

Yates's continuity correction

This is a correction that can improve the normal approximation to the distribution and result in a more accurate p-value with small sample sizes.

```
prop.test(cbind(y,n-y),correct=T)
```

```
    2-sample test for equality of proportions with
    continuity correction

data:  cbind(y, n - y)
X-squared = 1.079, df = 1, p-value = 0.2989
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01183885  0.04041028
sample estimates:
    prop 1     prop 2
0.10000000 0.08571429
```

The correction can be applied to the chi-squared test as well.

An alternative to this correction is to apply Fisher's exact test which gives an exact p-value.

Fisher's Exact Test

```
chisq.test(cbind(y,n-y),simulate.p.value=TRUE)
```

```
    Pearson's Chi-squared test with simulated
    p-value (based on 2000 replicates)

data:  cbind(y, n - y)
X-squared = 1.243, df = NA, p-value = 0.2994
```

We will describe how this test works later in the quarter in the context of permutation and resampling procedures.

In many practical situations all of the different tests described above give very similar results. By convention the chi-squared test is the most commonly used.