Methods of Mathematical Statistics

Notes by Tim Brown and Davide Ferrari

Module 8: Hypothesis Tests

Contents

1 Introduction and Tests About Proportions - 8.3

1.1 Introduction - Types of Error

Hypothesis Testing - Example

Parameter and interval estimation is the essential background for examining a specific hypothesis.

Science proceeds by changing theories when data shows that the current theory is inadequate.

This method extends outside science - data might show that a new procedure in a factory reduces the number of faulty devices.

Suppose currently a proportion $p = 0.06$ of circuits fail.

A new procedure is implemented and the number, Y, that fail in $n = 200$ circuits is observed.

Might decide procedure is an improvement if $Y \leq 7$ or $Y/n \leq 0.035$.

How sensible is this?

Type I Error

What could go wrong with this decision rule?

The new procedure may *not* reduce faults but by chance we observe at most 7 failures.

Then we would conclude the new procedure is better when it is not.

This is called a **Type I error**.

This error could be quite costly - changing a production line without reducing faults would be expensive.

Controlling the probability of **Type I error** will help manage this risk.

Type II Error

Could anything else could go wrong if **Type I error** is managed?

The new procedure *might* reduce faults but more than 7 failures are observed by chance.

Then the new procedure would be wrongly rejected.

This is called a **Type II error**.

This error would be less costly in the short term but might be much more costly long-term.

So whilst **Type I error** is often the one that is specifically controlled, **Type II error** remains important.

We manage **Type I error** and seek tests that minimise **Type II error**.

1.2 Null and Alternative Hypotheses

Describing Hypotheses

In view of this classification of errors, how should hypotheses be described?

Let p denote the proportion of faulty circuits produced by the factory using the new procedure.

No change hypothesis, or - the **Null Hypothesis** - in this example would be:

$$H_0 : p = 0.06.$$

The **Alternative Hypothesis** is what is accepted if evidence suggests the null hypothesis is wrong - here:

$$H_1 : p < 0.06.$$

The null hypothesis is **simple** because only one value is specified.

The alternative is **composite** because many values are specified.

Null Hypothesis Value - Least Favourable to Alternative

Null hypotheses are nearly always stated as simple hypotheses.

Null hypothesis $p = 0.06$ is on the boundary of the region $p \geq 0.06$ and is the "least favourable" element for the alternative hypothesis: its harder to differentiate between $p = 0.03$ and $p = 0.06$ than between $p = 0.03$ and $p = 0.5$.

Usually the null hypothesis is chosen to be the simple hypothesis that is on the boundary in this sense.

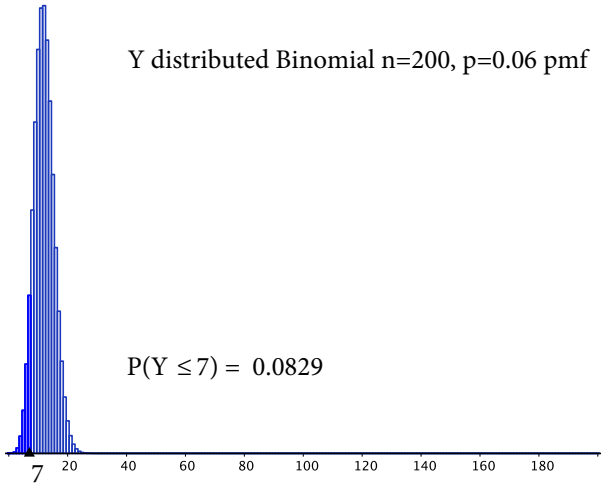

Figure 1: Significance Level is $P(Y \leq 7)$

1.3 Significance Level, Power

Significance Level

Type I error: Reject H_0 when H_0 is true.

Type II error: Fail to reject H_0 when H_1 is true.

The *significance level*, α, of the test is the probability of a Type I error.

In the example, since if $p = 0.06$, then $Y \sim Bin(200, 0.06)$ and from Figure 1

$$\alpha = P(Y \leq 7 | p = 0.06) = 0.0829$$

is the *significance level* of the test.

For $p > 0.06$ this prob. is smaller so $\sup_{p \geq 0.06} P(Y \leq 7|p) = 0.0829$.

Power

Suppose the new procedures actually worked and reduced the proportion to 0.03.

Then the probability of a Type II error, β, is shown in Figure 2 to be

$$\beta = P(Y > 7 | p = 0.03) = 0.254.$$

Halved the proportion of faulty circuits and still have a 25% chance of not rejecting null.

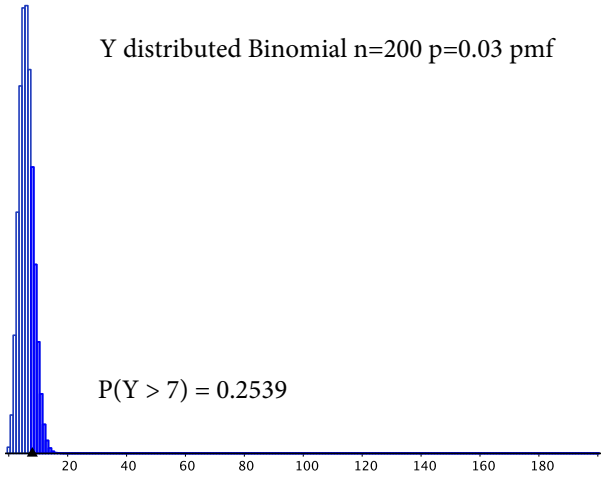

Figure 2: Type II Error is 0.2539

The *Power function of a test, K* , is the probability of rejecting H_0 when H_1 is true as a function of the parameter values in H_1 i.e. *power* is 1 - Type II error = probability of *correctly* rejecting null hypothesis.

Have shown $K(0.03) = 0.746$ and Figure 3 shows a plot of the power funtion vs. p.

Power Ctd.

As might be expected, the test is good at detecting values of p that are close to zero but not so good when p is close to 0.06.

Later - how to construct tests with good power at a given significance level.

Now - describe commonly used tests of a given significance level.

In practice, usually use the significance level $\alpha = 0.05$ - can live with incorrectly failing to reject null hypothesis 1 time in 20.

The 5% error rate is connected to the commonly used 95% for confidence intervals.

However, $\alpha = 0.01$ is not uncommon to offer more protection against Type I error.

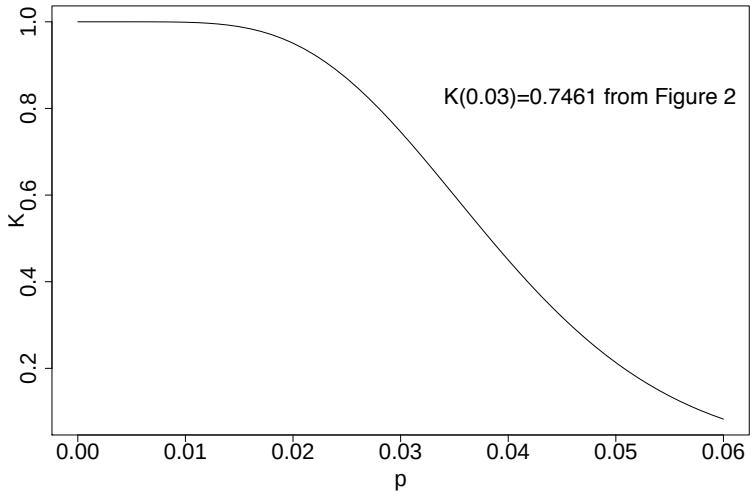

Figure 3: Plot of the power, K, as a function of p

1.4 Single Proportion

n Bernoulli trials success probability p

Suppose $Y \sim \text{Binomial}(n, p)$ - number of successes in Bernoulli trials

Test $H_0 : p = p_0$, $H_1 : p > p_0$, and take $\alpha = 0.05$.

Makes sense to reject H_0 if the observed value of Y is too large - ie $Y \geq c$ for some c.

$Y \geq c$ is often called the **critical region** of the test.

How do we choose c?

Need $P(Y \geq c | p = p_0) = \alpha$ for test to have significance level α, so c should be chosen to be the $100(1-\alpha)$ percentile of the Binomial(n, p_0) distribution.

For large n, when H_0 is true

$$Z = \frac{Y - np_0}{\sqrt{np_0(1-p_0)}} \approx N(0,1).$$

n Bernoulli trials success probability p Ctd

So if

$$c = np_0 + z_\alpha \sqrt{np_0(1-p_0)},$$

then

$$P(Y \geq c) = P\left(\frac{Y - np_0}{\sqrt{np_0(1-p_0)}} \geq z_\alpha \right) \approx \alpha.$$

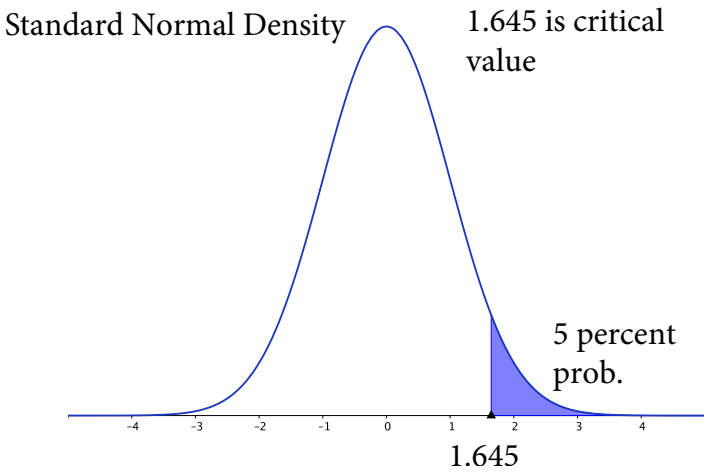

Figure 4: Critical Value for z for $>$ Alternative

Example 8.3-1. Commercial Dice -p is probability of rolling a six, Y number of sixes. $H_0 : p = 1/6$, $H_1 : p > 1/6$, $n = 8000$, so can use normal approximation.

Since $z_{0.05} = 1.645$, (see Figure 4)

$$c = 8000/6 + 1.645\sqrt{8000(1/6)(5/6)} = 1388.162$$

Observed $Y = 1389$ so reject H_0 at 5% level of significance and conclude that the die comes up with 6 too often.

Critical Value

In example, can equivalently use Z expressed as the standardised proportion of 6's,

$$Z = \frac{Y/n - 1/6}{\sqrt{5/(36n)}} \approx N(0,1).$$

Reject H_0 at level α if $Z > z_\alpha$: z_α is the *critical* value for Z.

In previous example,

$$z = \frac{1389/8000 - 1/6}{\sqrt{(1/6)(5/6)/8000}} = 1.67$$

and as the critical value is $z > z_{0.05} = 1.645$ we reject H_0.

Hypothesis Testing.

Have been conducting a one sided test because the *alternative is one-sided*, but the alternative hypothesis can be two-sided

$$H_0 : p = p_0, \text{ versus } H_1 : p \neq p_0.$$

Still compute the test statistic

$$Z = \frac{Y/n - p_0}{\sqrt{p_0(1-p_0)/n}} \approx N(0,1)$$

But now reject H_0 at level α if $|Z| > z_{\alpha/2}$.

Note that tests using the exact Binomial distribution will not generally have significance levels that are exactly a specified α, so usually a conservative approach is adopted and it is ensured that the significance level is below α.

1.5 Summary of Tests for Single Proportion

Summary of Critical Regions - Single Proportion

Can have one- or two-sided alternatives:

H_0	H_1	Critical Region
$p = p_0$	$p > p_0$	$z = \frac{y/n - p_0}{\sqrt{p_0(1-p_0)/n}} \geq z_\alpha$
$p = p_0$	$p < p_0$	$z = \frac{y/n - p_0}{\sqrt{p_0(1-p_0)/n}} \leq -z_\alpha$
$p = p_0$	$p \neq p_0$	$\lvert z\rvert = \frac{\lvert y/n - p_0\rvert}{\sqrt{p_0(1-p_0)/n}} \geq z_{\alpha/2}$

1.6 Example - Order of adding Milk to Tea (RA Fisher)

Example - Milk to Tea

A woman claims she can tell whether the tea or milk was added first when pouring tea.

Given 40 cups of tea, the order of adding milk was determined by tossing a coin and not revealed to the woman - called *randomised and blind* testing.

Woman was correct 29 times out of 40.

Is this evidence at the 5% level of significance that her claim was valid?

Example - Milk to Tea Ctd

Let p be the probability the woman gets the correct order in a single tasting. Hypotheses are:

$$H_0 : p = 0.5 \text{ (guessing) vs. } H_1 : p > 0.5 \text{ (claim valid).}$$

Need evidence *against* the hypothesis she is guessing.

One- sided alternative is appropriate because the claim is that the woman *can* tell the order.

Data $y/n = 29/40 = 0.725$

$$z = \frac{0.725 - 0.5}{\sqrt{.5 \times .5/40}} = 2.84$$

Critical value $z_{0.05} = 1.645$ (see Figure 4), therefore reject H_0.

Conclude that the data supports the woman's claim.

1.7 P-value

Hypothesis Testing: P-value

Common to report the results of a hypothesis test as a *p-value* - this is the *null hypothesis* probability of *data that supports the alternative as much or more than that observed.*

The null hypothesis is rejected if the p-value is less than the significance level.

In the previous example, large values of Z support the alternative hypothesis.

So the p-value from tables or R or Mathematica is given, for $Z \sim N(0, 1)$, by

$$P(Z > 2.846) = 0.00221.$$

As the p-value $0.00221 < 0.05$ we reject H_0 at the 5% level of significance.

Never *accept* a null hypothesis - instead if p-value > significance level, conclude there is *not enough evidence to reject* H_0 .

1.8 Example - Tea Testing R Approximate and Exact

Here is the R output with the approximate p-value illustrated in Figure 5.

```
prop.test(29, 40, p = 0.5, alternative = c("greater"),
    correct = F)

##
##  1-sample proportions test without
##  continuity correction
##
```

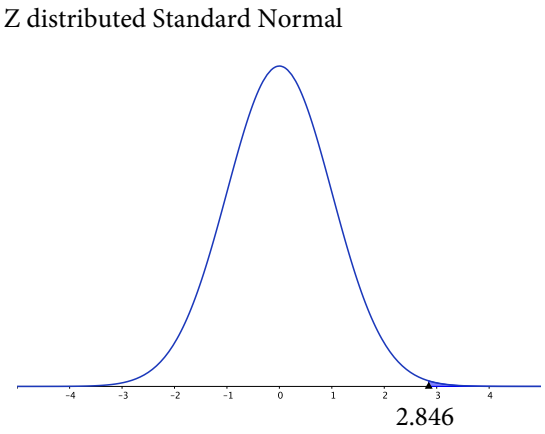

Figure 5: Approximate P-value is $P(Z > 2.846) = 0.00221$

```
## data:  29 out of 40, null probability 0.5
## X-squared = 8.1, df = 1, p-value =
## 0.002213
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.597457 1.000000
## sample estimates:
##     p
## 0.725
```

There is also an exact test based on the binomial probabilities with the exact p-value illustrated in Figure 6:

```
binom.test(29, 40, p = 0.5, alternative = c("greater"))

##
##  Exact binomial test
##
## data:  29 and 40
## number of successes = 29, number of
## trials = 40, p-value = 0.003213
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.5861226 1.0000000
## sample estimates:
## probability of success
##                  0.725
```

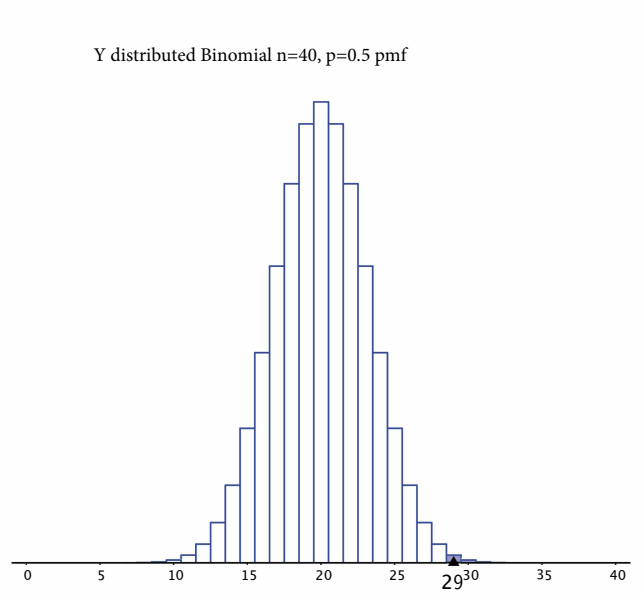

Figure 6: Exact P-value is $P(Y \geq 29) = 0.0.003213$

1.9 Tests about Two Proportions

Hypothesis Testing: Common Tests.

Comparing two proportions: p_1 and p_2 are the probabilities of success in two different populations.

Wish to test

$$H_0 : p_1 = p_2 \text{ versus } H_1 : p_1 > p_2.$$

based on independent samples (from the two populations) of size n_1 and n_2 with Y_1 and Y_2 successes.

Know

$$Z = \frac{Y_1/n_1 - Y_2/n_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n2}} \approx N(0,1).$$

Let $\hat{p}_1 = y_1/n_1$, $\hat{p}_2 = y_2/n_2$, $\hat{p} = (y_1 + y_2)/(n_1 + n_2)$. *Reject* H_0 at level α if

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n2)}} > z_\alpha.$$

1.10 Example - Insecticides

Experimental and Standard Insecticide

Two insecticides. Standard one kills 425 out of 500 mosquitoes, experimental one kills 459 out of 500. Is the experimental insecticide more effective?

Let $p_1(p_2)$ be the proportion of all mosquitoes killed by experimental (resp. standard) spray.)
The null and alternative hypotheses are:

$$H_0 : p_1 = p_2, \ H_1 : p_1 > p_2$$

The next two slides give the R output: in the first slide the calculations are done from the formulae.
In the second, the R command "prop.test" is used: this computes $P(Z^2 > 3.35760^2) = P(\chi^2(1) > 11.2732)$.
Figure 7 illustrates the p-value.

```
x <- c(459, 425)
n <- c(500, 500)
phat <- (x[1] + x[2])/(n[1] + n[2])
p1 <- x[1]/n[1]
p2 <- x[2]/n[2]
z <- (p1 - p2)/sqrt(phat * (1 - phat) * (1/n[1] +
    1/n[2]))
pvalue <- 1 - pnorm(z)
print(c(p1, p2, z, pvalue), digits = 3)

## [1] 0.918000 0.850000 3.357560 0.000393
```

```
prop.test(x, n, alternative = c("greater"), correct = FALSE)

##
##  2-sample test for equality of
##  proportions without continuity
##  correction
##
## data:  x out of n
## X-squared = 11.273, df = 1, p-value =
## 0.0003932
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.03487541 1.00000000
## sample estimates:
## prop 1 prop 2
##  0.918  0.850

# with continuity correction
prop.test(x, n, alternative = c("greater"))$p.value

## [1] 0.0005594062
```

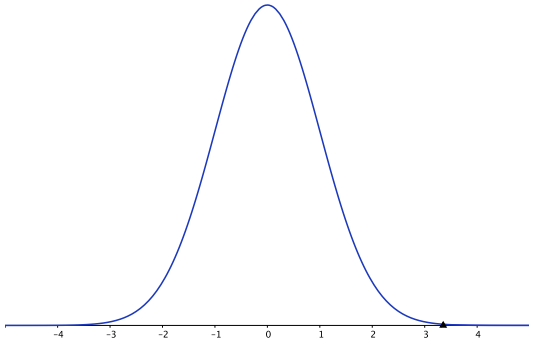

Figure 7: Approximate P-value is $P(Z > 3.358) = 0.000393$

1.11 Summary of Tests for Two Proportions

Hypothesis Testing: Common Tests.

As before can have one or two sided alternatives:

H_0	H_1	Critical Region
$H_0 : p_1 = p_2$	$H_1 : p_1 > p_2$	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n2)}} > z_\alpha$
$H_0 : p_1 = p_2$	$H_1 : p_1 < p_2$	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n2)}} < z_\alpha$
$H_0 : p_1 = p_2$	$H_1 : p_1 \neq p_2$	$\lvert z\rvert = \frac{\lvert\hat{p}_1 - \hat{p}_2\rvert}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n2)}} > z_{\alpha/2}$

2 Tests about One Mean - 8.1

2.1 Example - Tyres, Known σ

Example - Testing Tyres, Known σ

Next tests about one population mean enter.

Example Tyre manufacturer advertises that a new tyre will last 48,000km on average. A consumer group tests a sample of 50 tyres and finds the mean is 45,286km. Suppose the standard deviation is known to be $\sigma = 6012.6$km. Is this evidence against the manufacturers claim?

Let μ be the mean tyre life

$$H_0 : \mu = 48,000 \text{ versus } H_1 : \mu < 48,000.$$

Need evidence *against* the manufacturer to validly dispute the advertisement.

Example Tyres Ctd

Recall from the central limit theorem, whatver the distribution of the X's,

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

So reject $H_0 : \mu = \mu_0$ in favour of $H_1 : \mu < \mu_0$ at level α if

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{50}} < -z_\alpha.$$

Now $-z_{0.05} = -1.645$ (see Figure 4) and n is large so we calculate

$$z = \frac{45286 - 48000}{6021.6/\sqrt{50}} = -3.187.$$

So we reject H_0 at the 5% level of significance - $-3.187 < -1.96$ - and conclude the tyre life is lower than the advertised 48,000 km.

2.2 Summary of Tests for Single Mean, σ known

Critical Regions - Single Mean, σ known

Table gives critical regions for one- and two-sided tests for a population mean:

H_0	H_1	Critical Region
$\mu = \mu_0$	$\mu > \mu_0$	$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$ or $\bar{x} \geq \mu_0 + z_\alpha\sigma/\sqrt{n}$
$\mu = \mu_0$	$\mu < \mu_0$	$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha$ or $\bar{x} \leq \mu_0 - z_\alpha\sigma/\sqrt{n}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert z\rvert = \frac{\lvert\bar{x}-\mu_0\rvert}{\sigma/\sqrt{n}} \geq z_{\alpha/2}$ or $\lvert\bar{x} - \mu_0\rvert \geq z_{\alpha/2}\sigma/\sqrt{n}$

Note that the null hypothesis is *rejected* at significance level α, if, and only if, the value μ_0 is *outside* the one- or two-sided $100(1-\alpha)\%$ confidence interval since, for example,

$$\bar{x} \geq \mu_0 + z_\alpha\sigma/\sqrt{n} \Leftrightarrow \mu_0 \leq \bar{x} - z_\alpha\sigma/\sqrt{n}.$$

2.3 T- test - Unknown σ

Variance not known, small sample size

Often the variance is not known and the sample size is small.

If the sample is from a $N(\mu, \sigma^2)$ population with sample mean and sd $\bar{X}, S$, then the famous t-test uses T defined by:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

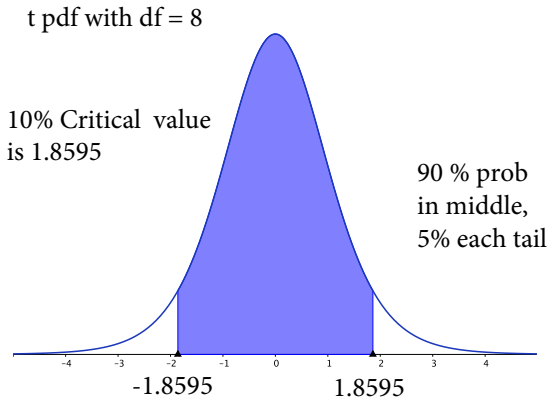

Figure 8: 10 % Critical Value for T-Test 8 df

Example $X \sim N(\mu, \sigma^2)$ is the growth (mm) in a tumor in a mouse.

$$H_0 : \mu = 4.0 \text{ versus } H_1 : \mu \neq 4.0$$

$n = 9$, significance level $\alpha = 0.1$.

Reject H_0 if

$$|t| = \frac{|\bar{x} - 4|}{s/\sqrt{9}} > t_{0.05}(8)$$

Example - Mice tumour growth

Conduct experiment with results: $\bar{x} = 4.3$, $s = 1.2$, $t_{0.05}(8) = 1.86$ (Figure 8).

$$t = \frac{4.3 - 4.0}{1.2/\sqrt{9}} = 0.75 < 1.86.$$

At the 10% level of significance we cannot reject H_0 and conclude there is not enough evidence that the tumour mean departs from 4mm.

Since t-values smaller or larger than 0.75 support the alternative hypothesis at least as much as our data, the p-value is

$$P(|T| \geq 0.75) = 2P(T \geq 0.75) = 0.475 > 0.05.$$

(In R , `2*(1-pt(0.75,8))` yields 0.4747312 (Figure 9).)

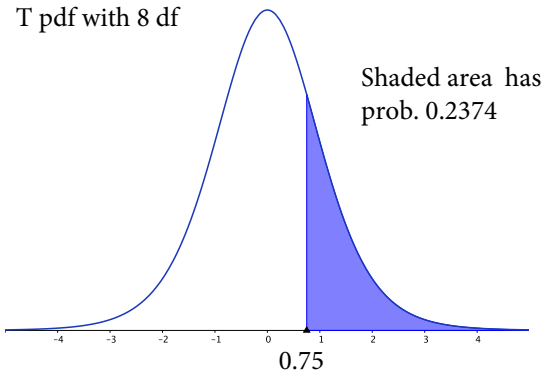

Figure 9: P-value is $2P(T > 0.75) = 0.475$

2.4 Summary of Tests for Single Mean, σ unknown

Summary of Critical Regions - Single Mean, σ unknown

H_0	H_1	Critical Region
$\mu = \mu_0$	$\mu > \mu_0$	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \geq t_\alpha$ or $\bar{x} \geq \mu_0 + t_\alpha s/\sqrt{n}$
$\mu = \mu_0$	$\mu < \mu_0$	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \leq -t_\alpha$ or $\bar{x} \leq \mu_0 - t_\alpha s/\sqrt{n}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert t\rvert = \frac{\lvert\bar{x}-\mu_0\rvert}{s/\sqrt{n}} \geq t_{\alpha/2}$ or $\lvert\bar{x}-\mu_0\rvert \geq t_{\alpha/2} s/\sqrt{n}$

where $t_\alpha = t_\alpha(n-1), t_{\alpha/2} = t_{\alpha/2}(n-1)$.

Note that the null hypothesis is *rejected* at significance level α, if, and only if, the value μ_0 is *outside* the one- or two-sided $100(1-\alpha)\%$ confidence interval since, for example,

$$\bar{x} \geq \mu_0 + t_\alpha(n-1)s/\sqrt{n} \Leftrightarrow \mu_0 \leq \bar{x} - t_\alpha(n-1)s/\sqrt{n}$$

2.5 Paired sample - t test

Example

As with confidence intervals, if there is one sample with two linked observations on each member of the sample, a one sample T statistic can be computed for the *differences*.

Tests using T are now for the equality, or otherwise, of the population means.

Next section considers *independent* samples for the same hypotheses of equality, or otherwise, of the population means.

3 Tests of the Equality of Two Means - 8.2

3.1 Example - Growth Hormone, One Sided

Example - Comparing Two Means Pooled Sample Variance

Suppose a botanist wants to compare effect of two different hormone concentrations on plant growth.

Rvs: X and Y represent growth in a random plant during the first 26 hours after treatment with hormone 1 & 2.

Suppose $X \sim N(\mu_X, \sigma^2)$ and $Y \sim N(\mu_Y, \sigma^2)$.

Expect less growth with hormone 1.

Then the appropriate null and alternative hypotheses are:

$$H_0 : \mu_X = \mu_Y, \text{ versus } H_1 : \mu_X - \mu_Y < 0.$$

Note that hypotheses should be decided before the data is collected.

Suppose plant samples of sizes n and m are gathered for X and Y.

Example - Growth Hormone, One Sided Ctd

Recall from Module 7, p.10

$$T = \frac{\bar{X} - \bar{Y}}{S_P\sqrt{1/n + 1/m}} \sim t(n + m - 2)$$

where S_P is the pooled variance estimate (a weighted average of the sample variances):

$$S_P = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n + m - 2}}.$$

This is a similar to the approach used for sample proportions and the same as for two sample confidence intervals.

Here it is assumed that the population variances are the same, & the samples are combined to estimate the common variance.

There no difference in the population probs was assumed, & samples combined to estimate the common population prob.

Example - Growth Hormone, One Sided Ctd 2

Reject H_0 if $t < -t_\alpha(n + m - 2)$.

Here $n = 11$, $m = 13$, $\bar{x} = 1.03$, $s_X^2 = 0.24$, $\bar{y} = 1.66$, $s_y^2 = 0.35$

$$S_P^2 = \frac{10 \times 0.24 + 12 \times 0.35}{11 + 13 - 2} = 0.3 = 0.548^2.$$

and hence

$$t = \frac{1.03 - 1.66}{\sqrt{0.3(1/11 + 1/13)}} = -2.81.$$

Critical t-value is $-t_{0.05}(22) = -1.717$ (see Figure 10) so we reject H_0 and conclude that there is statistically significant evidence of less growth with Hormone 1.

p-value (see Figure 11) is

$$P(T < -2.81) = 0.0051 < 0.05.$$

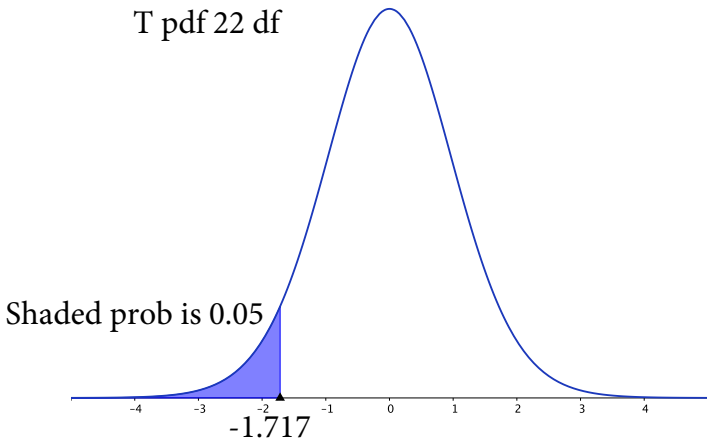

Figure 10: Critical 5% Value from t(22)

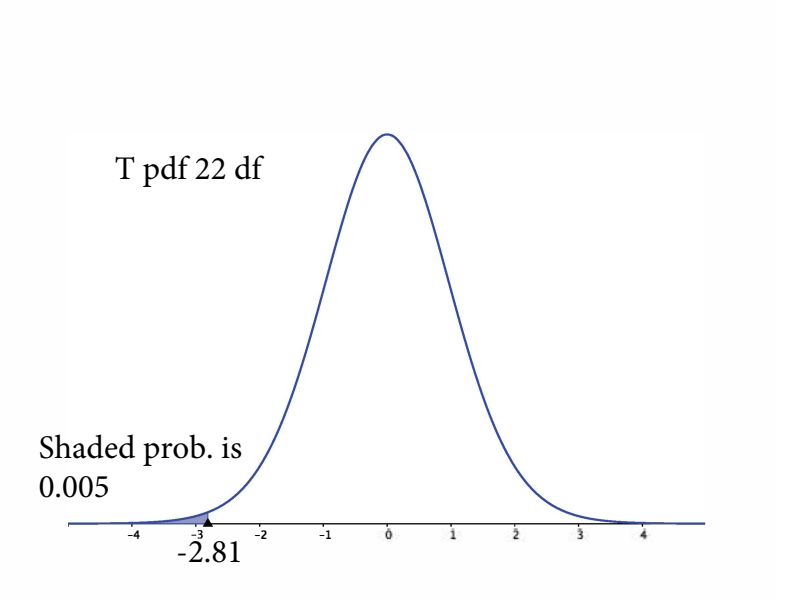

Figure 11: P value is $P(T < -2.81) = 0.005$

Rarely do it by hand: use `t.test` in R - "var.equal=T" gives pooled variance.

```
x = c(0.8, 1.8, 1, 0.1, 0.9, 1.7, 1, 1.4, 0.9,
    1.2, 0.5)
y = c(1, 0.8, 1.6, 2.6, 1.3, 1.1, 2.4, 1.8, 2.5,
    1.4, 1.9, 2, 1.2)
t.test(x, y, alternative = c("less"), var.equal = T)

##
##  Two Sample t-test
##
## data:  x and y
## t = -2.8112, df = 22, p-value = 0.005086
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.2468474
## sample estimates:
## mean of x mean of y
##  1.027273  1.661538
```

Boxplots of Data on Growth Hormones in Figure 12

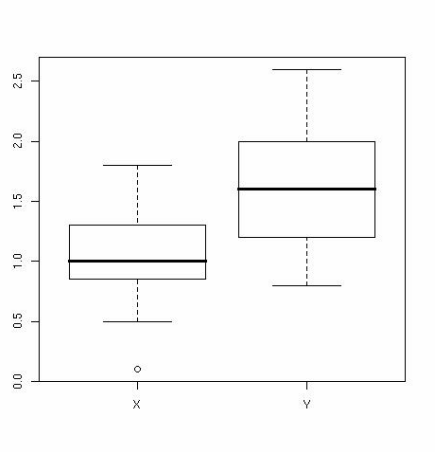

Figure 12: Boxplots for Hormones 1 and 2 - equal variance?

3.2 Example - Weights of Packages, Two Sided

Example - Weights of Packages

Weights in gm of packages, $X \sim N(\mu_X, \sigma^2)$ and $Y \sim N(\mu_Y, \sigma^2)$, filled by two methods (Example 8.2-2). Interested in testing:

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

Data and computer output (Figure 13) yield the p-value 0.052 so at the 5% level of significance there is not enough evidence to reject the null hypothesis of no difference in the means for the two methods.

Given closeness of result would probably collect more data!

Data and computer output:

```
x = c(1071, 1076, 1070, 1083, 1082, 1067, 1078,
    1080, 1075, 1084, 1075, 1080)
y = c(1074, 1069, 1075, 1067, 1068, 1079, 1082,
    1064, 1070, 1073, 1072, 1075)
t.test(x, y, var.equal = T)

##
##  Two Sample t-test
##
## data:  x and y
## t = 2.053, df = 22, p-value = 0.05215
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04488773  8.87822107
## sample estimates:
## mean of x mean of y
##  1076.750  1072.333
```

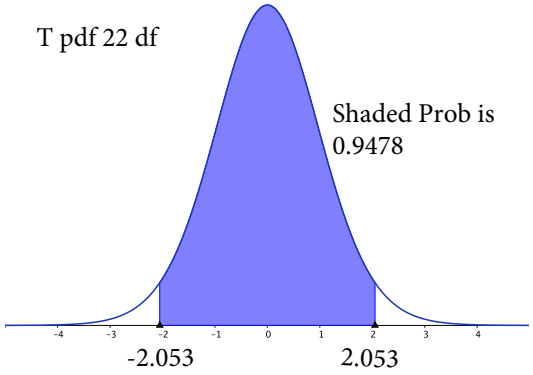

Figure 13: P value for Weights of Packages is $2P(T(22) > 2.053) = 0.052$

3.3 Example - Bubble Gum, One Sided, Welch/Pooled

Example - Bubble Gum Thickness, Cp. Pooled/Welch

$X \sim N(\mu_X, \sigma^2)$ and $Y \sim N(\mu_Y, \sigma^2)$ correspond to the thickness (hundredths of an inch) of regular gum and bubble gum with samples of sizes n=40 and m=50.

Text example 8.2-3. Because bubble gum has more elasticity than regular gum, the company wants to check it is being rolled out as thick as regular gum, so the appropriate hypotheses are:

$$H_0 : \mu_X = \mu_Y \text{ versus } H_1 : \mu_X < \mu_Y.$$

Data and output on the next pages.

Example - Bubble Gum, Pooled

```
data <- read.delim("Example_8_2-3.txt", header = T)
names(data)
```

```
## [1] "X"       "Y"       "Thiknes" "Code"
```

```
data[1:3, ]
```

```
##      X    Y Thiknes Code
## 1 6.85 7.10    6.85    X
## 2 6.60 7.05    6.60    X
## 3 6.70 6.70    6.70    X
```

```
z <- data[, 3]   #Thickness
w <- data[, 4]   # X or Y
```

Example - Bubble Gum, Pooled

```
t.test(z ~ w, var.equal = T)

##
##  Two Sample t-test
##
## data:  z by w
## t = -5.0524, df = 88, p-value =
## 2.345e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.19541537 -0.08508463
## sample estimates:
## mean in group X mean in group Y
##         6.70100         6.84125
```

Example - Bubble Gum, Welch

The boxplot in Figure 14 shows that the assumption of equal variances may not be valid.

Welch's t procedure was introduced in Module 7, p.14.

Recall that the degrees of freedom are lowered in exchange for not assuming a common variance.

So the p-value may be different with Welch's text.

Welch's test is the default in R and many other packages.

Here the conclusion on the next slide is similar in terms of the very small p-value.

So H_0 is clearly rejected and we conclude that the population mean thicknesses are different.

An alternative analysis (see text p.377) would use a normal approximation (OK with large sample sizes) with the standard error of the sample difference as $\sqrt{\frac{S_X^2}{50} + \frac{S_Y^2}{40}}$.

Example - Bubble Gum, Welch

```
t.test(z ~ w)

##
##  Welch Two Sample t-test
##
## data:  z by w
## t = -4.8604, df = 67.219, p-value =
## 7.357e-06
## alternative hypothesis: true difference in means is not equal to 0
```

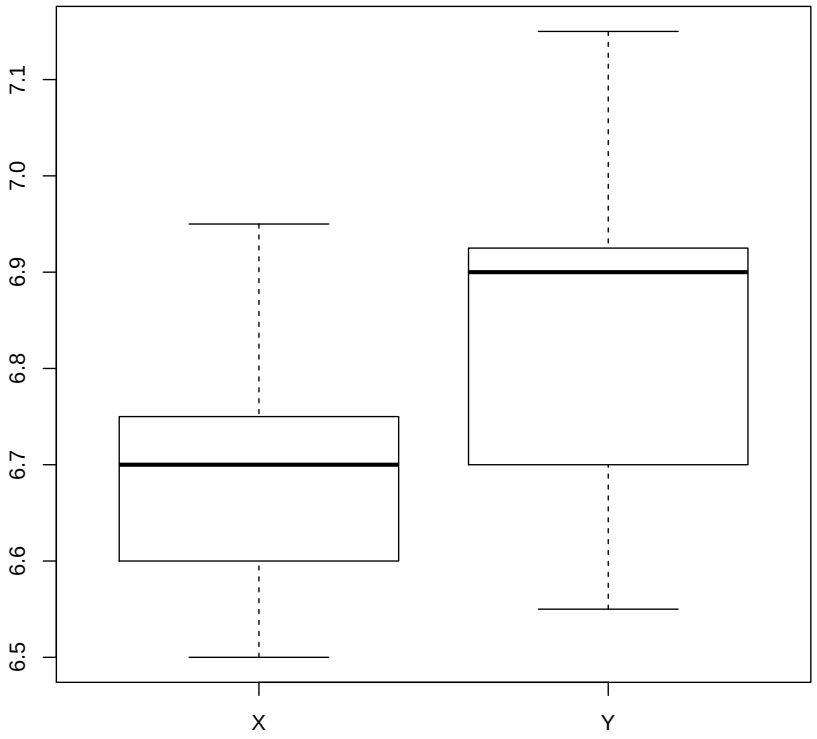

Figure 14: Boxplots of the thickness of gum and bubble gum

```
## 95 percent confidence interval:
##  -0.19784277 -0.08265723
## sample estimates:
## mean in group X mean in group Y
##         6.70100         6.84125
```

4 Distribution Free Tests - 8.4

4.1 Sign Test

Distribution Free Tests

Often don't want to assume normality but the sample is too small to appeal to the central limit theorem.

Can use nonparametric or distribution free methods - these don't make an assumption on an appropriate distribution for the data other than it is continuous.

In this case usually test hypotheses about the population median m. Eg.

$$H_0 : m = m_0 \text{ versus } H_1 : m \neq m_0.$$

The *sign test* is a hypothesis test based on same ideas as the confidence iterval for median.

Suppose $X_1, \ldots, X_n$ are a random sample from a continuous distribution.

Compute the number of negative signs in $X_1 - m_0, \ldots, X_n - m_0$.

Under H_0 this has a Binomial(n,1/2) distribution.

Example - Times between Calls

Times in seconds between calls to a switchboard (Example 8.4_1 in text). The null and alternative hypotheses are $H_0 : m = 6.2$ vs $H_1 : m < 6.2$ where m is the population median time between calls. The data and calculations are:

#	X	X-6.2	sign	#	X	X-6.2	sign
1	6.80	0.60	1	11	18.90	12.70	1
2	5.70	−0.50	−1	12	16.90	10.70	1
3	6.90	0.70	1	13	10.40	4.20	1
4	5.30	−0.90	−1	14	44.10	37.90	1
5	4.10	−2.10	−1	15	2.90	−3.30	−1
6	9.80	3.60	1	16	2.40	−3.80	−1
7	1.70	−4.50	−1	17	4.80	−1.40	−1
8	7.00	0.80	1	18	18.90	12.70	1
9	2.10	−4.10	−1	19	4.80	−1.40	−1
10	19.00	12.80	1	20	7.90	1.70	1

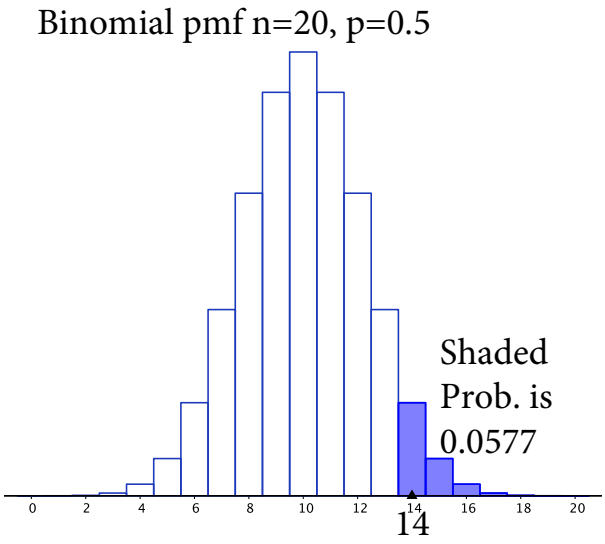

Figure 15: Critical Value for Sign Test n=20 is 14

Example - Times between Calls Ctd

Suppose Y is the number of negative signs. Reject H_0 if Y is too large. (median < 6.2 then expect more than $1/2$ of the observations < 6.2).

Have $P(Y \geq 14) = 1 - P(Y \leq 13) = 0.0577$ so 14 is approximate 5% critical value
(In R `> 1-pbinom(13,20,0.5)` Figure 15)

As observed $y = 9 < 14$, we do not reject H_0 .

p-value is $P(Y \geq 9) = 1 - P(Y \leq 8) = 0.75 > 0.05$, so again can't reject H_0. (In R `> 1-pbinom(8,20,0.5)`, see Figure 16.)

The sign test requires few assumptions.

But it doesn't use information on the size of the differences, so it can be insensitive to departures from H_0 ie large Type II error or small power.

4.2 Wilcoxon One Sample Signed Rank Test

Wilcoxon One Sample Signed Rank Test

Assume the underlying distribution is continuous and *symmetrical*.

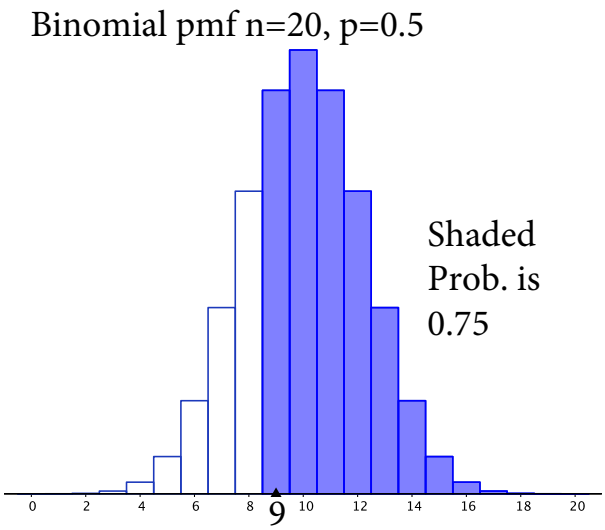

Figure 16: P-Value for Sign Test is $P(Y \geq 9) = 0.75$

Null hypothesis is again $H_0 : m = m_0$ against a one-sided or two-sided alternative.

Rank $|X_1 - m_0|, \ldots, |X_n - m_0|$.

Wilcoxon signed rank statistic W is the sum of the signs times the rank.

Warning: the literature is ambiguous about the definition of the test and some packages work out slightly different statisitics.

4.3 Example - Fish

Example - Fish

Lengths of 10 sunfish are 5.0, 3.9, 5.2, 5.5, 2.8, 6.1, 6.4, 2.6, 1.7, 4.3.

Null hypothesis is $H_0 : m = 3.7$ versus $H_1 : m > 3.7$.

Calculations:

	1	2	3	4	5
$x_i - m_0$	1.30	0.20	1.50	1.80	−0.90
$\lvert x_i - m_0 \rvert$	1.30	0.20	1.50	1.80	0.90
Ranks	5.00	1.00	6.00	7.00	3.00
Signed Ranks	5.00	1.00	6.00	7.00	−3.00
	6	7	8	9	10
$x_i - m_0$	2.40	2.70	−1.10	−2.00	0.60
$\lvert x_i - m_0 \rvert$	2.40	2.70	1.10	2.00	0.60
Ranks	9.00	10.00	4.00	8.00	2.00
Signed Ranks	9.00	10.00	−4.00	−8.00	2.00

Example - Fish Ctd

Hence $W = 5 + 1 + 6 + 7 - 3 + 9 + 10 - 4 - 8 + 2 = 25$.

Alternatively sum of positive ranks is $V = 5 + 1 + 6 + 7 + 9 + 10 + 2 = 40$.

Note $W = 2V - \frac{n(n+1)}{2}$.

What is an appropriate critical region? For simplicity assume no ties.

If $H_1 : m > 3.7$ is true expect more positive signs. Then W should be large, so critical region should be $\{w : w > c\}$ for suitable c.

If H_0 is true then $P(X_i < m_0) = P(X_i > m_0) = 1/2$.

Assignment of the n signs to the ranks are mutually independent .

Statistic W is the sum of the integers $1, \ldots, n$, each with a positive or negative sign.

Under H_0, $W \sim \sum_{i=1}^{n} W_i$ where

$$P(W_i = i) = P(W_i = -i) = \frac{1}{2}, \ i = 1, \ldots, n.$$

Example - Fish Ctd 2

Mean under H_0 is $E(W_i) = -i(\frac{1}{2}) + i(\frac{1}{2}) = 0$ so $E(W) = 0$.

Similarly, $Var(W_i) = E(W_i^2) = i^2$ and

$$\text{Var}(W) = \sum_{i=1}^{n} \text{Var}(W_i) = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

A more advanced argument shows that for large n when H_0 is true,

$$Z = \frac{W - 0}{\sqrt{n(n+1)(2n+1)/6}} \approx N(0, 1).$$

So $P(W \geq c|H_0) \approx P(Z > z_\alpha|H_0)$, giving approximate c.

For $H_0 : m = 3.7$ versus $H_1 : m > 3.7$, $n = 10$, $\alpha = 0.05$, critical region:

$$z = \frac{w}{\sqrt{10(11)(21/6}} \geq 1.645$$

(see Figure 4).

Example - Fish Ctd 3

Alternatively

$$w \geq 1.645\sqrt{\frac{10(11)(21)}{6}} = 32.27.$$

Hence not enough evidence to reject H_0 that the median is 3.7.

The R output illustrates the command `wilcox.test`, tailoring the output via storage and printing in a data frame.

Boxplot in Figure 17 shows no compelling evidence against the symmetry assumption.

Example - Fish Ctd 4, R Output

```
X <- c(5.0, 3.9, 5.2, 5.5, 2.8, 6.1, 6.4, 2.6,
1.7, 4.3)
out<- wilcox.test(X,mu=3.7,alternative=c("greater"),
exact=TRUE);
out$method

## [1] "Wilcoxon signed rank test"

data.frame(data=c(out$data.name),V=c(out$statistic),
p.value=c(out$p.value),row.names=c(""))

##  data  V   p.value
##     X 40 0.1162109
```

Example - Fish Ctd 5, R Output

Comments on the Wilcoxon Signed Rank Test

Other versions of the test use the sums of the positive ranks but give the same p-values (R gives the sum of positive ranks V).

Text uses moment generating functions to compute the exact distribution for a couple of small sample sizes - this is easily implemented in Mathematica (see Lab 12, question 6) for an arbitrary sample size.

Often applied to paired data. Eg.two measures for each a random sample of people.

4.4 Wilcoxon - Two Sample Test

Wilcoxon - Two Sample Test

Independent random samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ from 2 different populations with medians m_X and m_Y respectively.

Hypotheses: $H_0 : m_X = m_Y$ versus one or two-sided alternative.

Order the *combined* sample. Let W be the sum of the ranks of $Y_1, \ldots, Y_{n_2}$.

If $m_Y > m_X$ expect W to be large (as most of the big ranks will be for the Y's).

Can determine the H_0 mean and variance of W: $\mu_W = E(W) = \frac{n_2(n_1+n_2+1)}{2}$, $\sigma^2_W = Var(W) = \frac{n_1 n_2(n_1+n_2+1)}{12}$.

There is a normal approximation: $Z = \frac{W-\mu_W}{\sigma_W} \approx N(0,1)$ for large n_1, n_2.

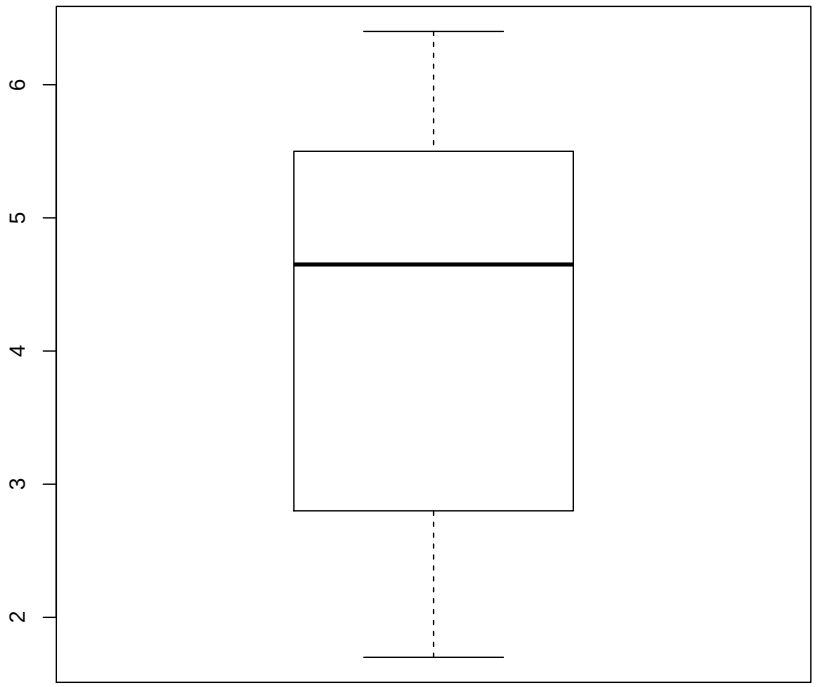

Figure 17: Boxplot of sunfish lengths

4.5 Example - Wilccoxon Two Sample, Packing Cinnamon

Example - Packing Cinammon

Two companies package cinnamon. It is thought that company X maybe underfilling the packages relative to company Y. Samples of size eight from each company yield the weights:

	1	2	3	4
X	117.1	121.3	127.8	121.9
Y	123.5	125.3	126.5	127.9
	5	6	7	8
X	117.4	124.5	119.5	115.1
Y	122.1	125.6	129.8	117.2

Test $H_0 : m_X = m_Y$ versus $H_1 : m_X < m_Y$.

Sorted data in the combined sample is:

115.1 (X)	117.1(X)	117.2(Y)	117.4(X)
119.5(X)	121.3(X)	121.9(X)	122.1(Y)
123.5(Y)	124.5(X)	125.3(Y)	125.6(Y)
126.5(Y)	127.8(X)	127.9(Y)	129.8(Y)

Example - Packing Cinammon Ctd

So the ranks for company Y are $3, 8, 9, 11, 12, 13, 15, 16$ and $w = 3+8+\cdots+16 = 87$.

The z-value for the data is:

$$z = \frac{87 - 8 * 17/2}{\sqrt{8 * 8 * 17/12}} = 1.995.$$

So the p-value from the R command:
`1-pnorm((87-4*17)/sqrt(64*17/12))`
is 0.023.

Hence reject H_0 and conclude at the 5% level of significance that the median weight for company X is less than for company Y.

As with the one-sample signed rank test, there are some disagreements in the literature about the most appropriate statistic.

Example - Packing Cinammon Ctd 2

As with the one-sample signed rank test, there are some disagreements in the literature about the most appropriate statistic.

R uses a slightly different definition of the statistic.

Unlike the signed rank statistic the p-value can be slightly different (and not just due to a continuity correction).

This is illustrated in the R output below where the p-value is 0.025 rather than 0.023.

The conclusion remains: reject the null hypothesis in favour of the alternative that the median weight for company X is less than for company Y.

Example - Packing Cinammon Ctd 3

```
X <- c(117.1, 121.30, 127.8, 121.9,
117.4 , 124.5 , 119.5 , 115.1)
Y <- c(123.5, 125.3, 126.5, 127.9,
122.1, 125.6, 129.8, 117.2)
out <- wilcox.test(X,Y,mu=0,alternative=c("less"))
out$method

## [1] "Wilcoxon rank sum test"

data.frame(data=c(out$data.name),W=c(out$statistic),
p.value=c(out$p.value),row.names=c(""))

##     data  W    p.value
##  X and Y 13 0.02494172
```

5 Chi Square Goodness of Fit - 9.1

5.1 Introduction, Binomial Model

Goodness of Fit Tests

Goodness of fit tests - how well does a given model fit a set of data?

So far this has been examined using diagnostic plots, but formal tests can be based on Pearson's chi-square statistic.

Eg: Binomial model $Y_1 \sim \text{Bin}(n, p_1)$. For large n,

$$Z = \frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}} \approx N(0,1),$$

Hence $Q_1 = Z^2 \approx \chi^2_1$ for large n.

Testing $H_0 : p = p_1$ versus $H_1 : p \neq p_1$: would reject H_0 if $|Z|$ and hence Q_1 were too large.

Introduction, Binomial Model 2

But

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1(1-p_1)} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_1 - np_1)^2}{n(1-p_1)}.$$

And

$$(Y_1 - np_1)^2 = (n - Y_1 - n(1-p_1))^2 = (Y_2 - np_2)^2$$

where $Y_2 = n - Y_1$ and $p_2 = 1 - p_1$.

Hence

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1(1-p_1)} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2}.$$

Introduction, Binomial Model 3

Y_1 is the *observed* number of successes, np_1 the *expected* number of successes.

Y_2 is the *observed* number of failures, np_2 the *expected* number of failures.

So

$$Q_1 = \sum_{i=1}^{2} \frac{(Y_i - np_i)^2}{np_i} = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} \approx \chi_1^2$$

where O_i is the observed number and E_i is the expected number.

Even though there are two cells, there is one degree of freedom, because the value of $Y_2 = n - Y_1$ is known from the value ofY_1 and the sample size n.

5.2 More than one outcome or category

More than one category

Generalize to k possible outcomes. p_i probability of ith outcome. $(\sum_{i=1}^{k} p_i = 1)$

Suppose n trials withY_i number of type i outcomes and np_i the expected number of type i.

The probability distribution of $(Y_1, \cdots, Y_k)$ is called *multinomial*.

Arguing as for the Binomial distribution: provided $y_i \in \{0, \cdots, n\}, \sum_{i=1}^{k} y_i = n$

$$P(Y_1 = y_1, \cdots, Y_k = y_k) = \frac{n!}{y_1! \cdots y_k!} p_1^{y_1} \cdots p_k^{y_k}$$

Now, as the sample size, n, gets large:

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \approx \chi_{k-1}^2$$

More than one category

Df are $k-1$ as $Y_k = n - Y_1 - \cdots - Y_{k-1}$.

Approximation is good if all $np_i \geq 5$.

If we have a model for the p_i, we can test the null hypothesis that this model generated the data using Q_1 against the alternative that the model could not have generated the data.

If Q_1 is small the model is reasonable, otherwise it is not.

5.3 Example - Transport

Example - Transport

Past records indicate that the proportions of commuters using various types of transport are:

Type	Bus	Train	Car	Other
Proportion	.25	.15	.50	.1

After a three month campaign, a random sample of 80 found:

Type	Bus	Train	Car	Other
Number	26	15	32	7

Is this evidence the campaign altered commuters behaviour?

Expected frequencies are:

Type	Bus	Train	Car	Other
Expected	20	12	40	8

and

$$\chi^2 = \frac{(26-20)^2}{20} + \frac{(15-12)^2}{12} + \frac{(32-40)^2}{40} + \frac{(7-8)^2}{8} = 4.275.$$

Example - Transport 2

Test H_0: Proportions same vs. H_1: Proportions different.

χ^2 above has a $\chi^2(3)$ distribution under the null hypothesis, because there are 4 types of public transport.

Now $\chi^2_{0.05}(3) = 7.81$ so there is not enough evidence the proportions have changed.

p-value (see Figure 18) is:

$$P(\chi^2(3) > 4.275) = 0.233 > 0.05.$$

Example - Transport 3, R output

```
x <- c(26, 15, 32, 7)
p <- c(0.25, 0.15, 0.5, 0.1)
(t1 <- chisq.test(x, p = p))

##
##  Chi-squared test for given
##  probabilities
##
## data:  x
## X-squared = 4.275, df = 3, p-value =
## 0.2333
```

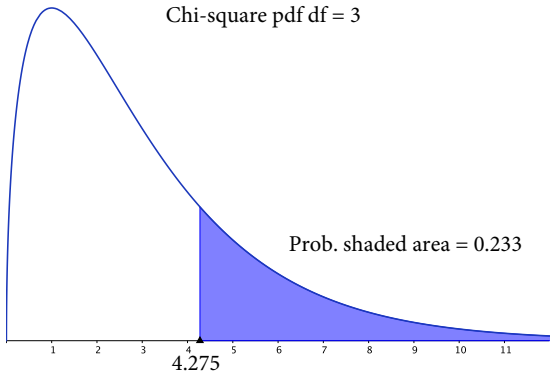

Figure 18: P-value for Chi-Square 3 is $P(\chi^2 \geq 4.275) = 0.233$

5.4 Estimating Parameters

Example - Alpha Particles

Can't always specify the exact model because it contains parameters that require estimation.

Suppose X is number of alpha particles emitted in 0.1 sec by a source.

Fifty observations are: 7, 4, 3, 6, 4, 4, 5, 3, 5, 3, 5, 5, 3, 2, 5, 4, 3, 3, 7, 6, 6, 4, 3 11, 9, 6, 7, 4, 5, 4, 7, 3, 2, 8, 6, 7, 4, 1, 9, 8, 4, 8, 9, 3, 9, 7, 7, 9, 3, 10.

Is the data consistent with a Poisson distribution?

Only specifying the family of the distribution, not the parameters.

Test $H_0 : X \sim$ Poisson versus $H_1 : X \sim$ Something else.

Estimate the Poisson parameter λ by MLE $\hat{\lambda} = \bar{x} = 5.4$.

Need to *reduce the degrees of freedom by the number of parameters estimated*, here by 1.

Does the Poisson(5.4) model give an adequate fit?

Example - Alpha Particles Ctd

Consider appropriate categories. (Collapse data)

```
X <-c(7,  4,  3,  6, 4, 4, 5, 3, 5, 3,
5, 5, 3, 2, 5, 4, 3, 3, 7, 6, 6, 4, 3,
11, 9, 6, 7, 4, 5, 4, 7, 3, 2, 8, 6,
7, 4, 1, 9, 8, 4, 8, 9, 3, 9, 7, 7, 9, 3, 10)
X1 <- cut(X,breaks=c(0,3.5,4.5,5.5,6.5,7.5,100))
(T1 <- table(X1))
```

```
## X1
##   (0,3.5] (3.5,4.5] (4.5,5.5] (5.5,6.5]
##        13         9         6         5
## (6.5,7.5] (7.5,100]
##         7        10
```

Example - Alpha Particles Ctd 2

Calculate appropriate Poisson probs.

```
n <- sum(X)
p1 <- sum(dpois(0:3, 5.4))
p2 <- dpois(4, 5.4)
p3 <- dpois(5, 5.4)
p4 <- dpois(6, 5.4)
p5 <- dpois(7, 5.4)
p6 <- 1 - (p1 + p2 + p3 + p4 + p5)
p <- c(p1, p2, p3, p4, p5, p6)
```

Example - Alpha Particles Ctd 3

Carry out test:

```
chisq.test(as.vector(T1), p = p)

##
##  Chi-squared test for given
##  probabilities
##
## data:  as.vector(T1)
## X-squared = 2.7334, df = 5, p-value =
## 0.741

# Wrong deg. freedom - reduce by 1:
1 - pchisq(2.7334, 4)

## [1] 0.6033828
```

Example - Alpha Particles Ctd 4

Need to adjust p-values as we have estimated the mean.

Critical value is $\chi^2_{0.05}(4) = 9.488$ so we cannot reject H_0.

Not enough evidence against the Poisson model (Figure 19) - good fit as table shows:

	0:3	4	5	6	7	8+
Observed	13.00	9.00	6.00	5.00	7.00	10.00
Expected	10.66	8.00	8.64	7.78	6.00	8.92

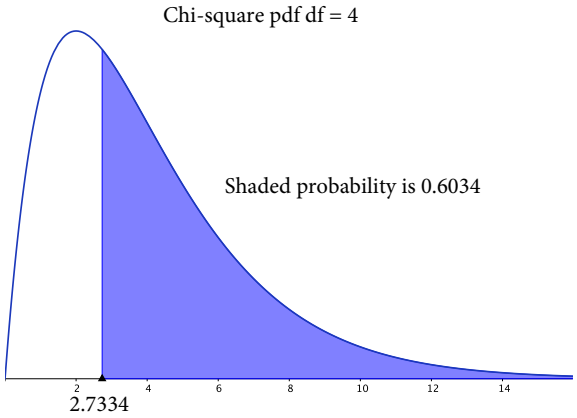

Figure 19: P-value is $P(\chi^2 \geq 2.7334) = 0.603$

6 Contingency Tables - 9.2

6.1 Tests of Independence in Contingency Tables

Tests of Independence in Contingency Tables

Contingency Tables are tables recording the number of cases if a sample is classified in two or more ways.

Skip first part of 9.2 and consider tests of independence between classifications - starts p.436.

Two attributes associated with each outcome. $A_1, \ldots A_k$ or $B_1, \ldots, B_h$. Eg. A may be a height category and B a weight category.

$$p_{ij} = P(A_i \cap B_j), i = 1, \ldots, k, \ j = 1, \ldots, h$$

.

Wish to test the hypothesis

$$H_0 : p_{ij} = P(A_i)P(B_j), i = 1, \ldots, k, \ j = 1, \ldots, h \text{ versus}$$

$$H_1 : \text{not } H_0.$$

6.2 Example - Gender/Order in Family

Example - Gender/Order in Family

150 executives were classified by sex, A, and whether or not they were first born, B. Table is:

	First Born	Not	Total
Male	34	74	108
Female	20	22	42
Total	54	96	150

Test H_0 : among executives, sex and birth order are independent versus H_1 : not independent.

Example - Gender/Order in Family Ctd

Recall discrete bivariate distributions

	First Born	Not	
Male	p_{11}	p_{12}	$p_{1.}$
Female	p_{21}	p_{22}	$p_{2.}$
Total	$p_{.1}$	$p_{.2}$	1

Marginals are $p_{i.} = \sum_{j=1}^{h} p_{ij} = P(A_i)$ and $p_{.j} = \sum_{i=1}^{k} p_{ij} = P(B_j)$

Null hypothesis of independence is just, $H_0 : p_{ij} = p_{i.}p_{.j}$

		First Born	Not	
Data:	Male	y_{11}	y_{12}	$y_{1.}$
	Female	y_{21}	y_{22}	$y_{2.}$
	Total	$y_{.1}$	$y_{.2}$	n

Example - Gender/Order in Family Ctd 2

Estimates, $\hat{p}_{i.} = y_{i.}/n$, $\hat{p}_{.j} = y_{.j}/n$, where $y_{i.} = \sum_{j=1}^{h} y_{ij}$ and n is the overall sample size.

χ^2 goodness of fit statistic for given p_{ij} is

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{h} \frac{(Y_{ij} - np_{ij})^2}{np_{ij}}$$

Under H_0 an estimator of p_{ij} is

$$\hat{p}_{ij} = \hat{p}_{i.}\hat{p}_{.j} = \frac{Y_{i.}Y_{.j}}{n^2}$$

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{h} \frac{(Y_{ij} - Y_{i.}Y_{.j}/n)^2}{Y_{i.}Y_{.j}/n} \approx \chi^2(k-1)(h-1)$$

because estimate $k-1$ probs. for rows and $h-1$ for columns so df $= hk - 1 - (h-1) - (k-1) = (h-1)(k-1)$

Explanation for Degrees of Freedom

Rule of thumb: number of unknown observations - number of parameters estimated.

In contingency tables, number of unknown observations is number of categories minus 1.

Independence means row and column marginal distributions are estimated - each sums to 1 so number of paramters for each is number of categories minus 1.

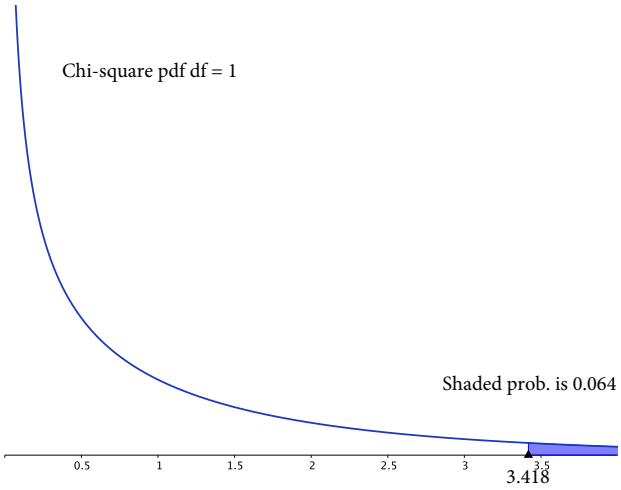

Figure 20: P-value is $P(\chi^2 \geq 3.418) = 0.064$

Example - Gender/Order in Family Ctd 3, R output

```
# note order
X <- matrix(c(34, 20, 74, 22), nrow = 2)
(c1 <- chisq.test(X, correct = F))

##
##  Pearson's Chi-squared test
##
## data:  X
## X-squared = 3.418, df = 1, p-value =
## 0.06449
```

So we do not have enough evidence to reject H_0 at the 5% level of significance (Figure 20) (expected is $y_{i.}y_{.j}/n$)

Other hypotheses

Sometimes the data collected is from different groups, with each group member being allocated to one of a number of categories.

This is covered in the text in 9.2 from 433 to 436.

The null hypothesis becomes that the different groups have the same multinomial distribution for the numbers in the group in the categories.

Remarkably, the chi-square statistic is computed the same way, and the degrees of freedom calculated the same way.

Lab and Workshop 12 Questions 14 and 15 consider two examples.

7 Likelihood Ratio Test

Likelihood Ratio Test.

- The likelihood ratio test is a general procedure that can be applied for example when both H_0 and H_1 are composite.
- Probably the most common method of generating tests in practice.
- Parameter space Ω, and $\omega \subset \Omega$.
- $H_0 : \theta \in \omega$ versus $H_1 : \theta \in \omega^c$.
- Likelihood ratio:
$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})}$$
where $L(\hat{\omega})$ is the maximum of the likelihood when H_0 is true and $L(\hat{\Omega})$ is the maximum over all of Ω.

Likelihood Ratio Test.

- $$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})}$$
- Clearly $0 \leq \lambda \leq 1$.
- λ near zero then the data do not support $H_0 : \theta \in \omega$
- λ near one then the data do support $H_0 : \theta \in \omega$
- The critical region is thus given by the set of sample points for which
$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} \leq k$$
$0 < k < 1$ and k is chosen to give the desired significance level.

Likelihood Ratio Test.

- Example 9.3-1. $X \sim N(\mu, 5)$. $H_0 : \mu = 162$ versus $H_1 : \mu \neq 162$
- When H_0 is true, $\mu = 162$ so $L(\hat{\omega}) = L(162)$.
- Recall $\hat{\mu} = \bar{x}$ is the MLE so $L(\hat{\Omega}) = L(\bar{x})$.
- Rearrangement yields
$$\lambda = \frac{(10\pi)^{-n/2} \exp\left[-\frac{1}{10}\sum_{i=1}^{n}(x_i - 162)^2\right]}{(10\pi)^{-n/2} \exp\left[-\frac{1}{10}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]} = \exp[-\frac{n}{10}(\bar{x} - 162)^2]$$
- So $\bar{x}$ near 162 supports H_0 etc.

Likelihood Ratio Test.

- $\lambda \leq k$ same as
$$\frac{|\bar{x} - 162|}{\sigma/\sqrt{n}} \geq c$$
- Critical region for a size α test is
$$C = \{\bar{x} : \frac{|\bar{x} - 162|}{\sigma/\sqrt{n}} \geq z_{\alpha/2}\}$$
- This required knowledge of the distribution of $\bar{x}$!

Likelihood Ratio Test.

- $X_1, \ldots, X_n$ random sample from $N(\mu, \sigma^2)$, $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.
- $$\omega = \{(\mu, \sigma^2) : \mu = \mu_0, 0 < \sigma^2 < \infty\}$$
and $\hat{\mu} = \mu_0$, $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (x_i - \mu_0)^2$.
- $$\Omega = \{(\mu, \sigma^2) : -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$$
and $\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$.
- Some simplification yields
$$\lambda = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \mu_0)^2} \right]^{n/2}$$

Likelihood Ratio Test.

- and the Analysis of variance identity is:
$$\sum_{i=1}^n (x_i - \mu_0)^2 = \sum_{i=1}^n (x_i - \bar{x} + \bar{x}\mu_0)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2.$$
- Substitute and rearrange to get
$$\lambda = \left[\frac{1}{1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right]^{n/2}.$$

Likelihood Ratio Test.

- This is $\leq k$ when
$$\frac{n(\bar{x}-\mu_0)^2}{\sum_{i=1}^n (x_i-\bar{x})^2/(n-1)} \geq k_1$$
- When H_0 is true $\sqrt{n}(\bar{X}-\mu_0)/\sigma \sim N(0,1)$ and $\sum_{i=1}^n (X_i-\bar{X})^2/\sigma^2 \sim \chi^2(n-1)$, and is independent of $\bar{X}$.
- But
$$T = \frac{\sqrt{n}(\bar{X}-\mu_0)/\sigma}{\sqrt{\sum_{i=1}^n (X_i-\bar{X})^2/\{(n-1)\sigma^2\}}} = \frac{\sqrt{n}(\bar{X}-\mu_0)}{\sqrt{\sum_{i=1}^n (X_i-\bar{X})^2/(n-1)}} = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t(n-1)$$
- So we reject H_0 when T^2 is too large or level α critical region is
$$|T| \geq t_{\alpha/2}(n-1)$$

Likelihood Ratio Test.

- Usually easy to find the form of the test.
- Manipulating until we have something whose distribution we know can be tricky!
- Many of the standard tests arise form the likelihood ratio.
- In MAST90104 there will be many examples of likelihood ratio tests and their asymptotic distributions will often be used.