Hypothesis Testing[1]

November 10, 2017

[1] HMS, 2017, v1.4

Chapter References

- Diez: Chapter 4.3, 4.5,
- Navidi, Chapter 6.0, 6.1, 6.2, (self read 6.3), 6.4, 6.5, (self read 6.6), 6.7, 6.10, 6.11

Inferential Statistics

- Using sample data to infer something about the population.

Terminology

- **Confidence Intervals:** Estimating a population parameter.
- **Tests of significance:** To assess the evidence provided by data about some claim on the population.
  1. A formal procedure for comparing observed data with a claim (called a hypothesis) whose truth we want to access.
  2. The hypothesis is a statement about a parameter, such as a mean.
  3. We express the results of a significance test in terms of a probability that measures how well the data and the hypothesis agree.

Two Types of Hypotheses

There are two types of hypotheses:

- **The Null Hypothesis** states that there is **no difference** between a parameter and a specific value. The Null Hypothesis is denoted by the symbol H_o and is what we want to disprove
- **The Alternative Hypothesis** - states that there **is a difference** between a parameter and a specific value. The alternative hypothesis is denoted by the symbol H_1

Examples of H_o

- Fourth graders at a school perform equally well in math compared to fourth graders at another school
- Babies born in the US are on average the same weight at birth compared to babies born in the UK.
- Two groups of nematode worms are treated differently, both sets of worm appear to have the same life span.
- Artists are no more likely to be left-handed than people in the general population.
- On average, the dose of aspirin in a single tablet is 200 mg
- Women and men are equally likely to be vegetarian
- People are likely to loose weight whether they are on a protein or carbohydrate diet.
- The percentage tip left at a family or fine dinning restaurant is the same.
- Age has no effect on mathematical ability
- There is no difference in pain relief after chewing willow bark versus taking a placebo

H_o

The null hypothesis says that what ever difference you observed is due to random chance alone. The effect is therefore not real.

In statistics, a hypothesis is a claim about some aspect of a population.

A hypothesis test allows us to test the claim about the population and find out how likely it is to be true

Symbolic Representation of H

When writing down H_o and H_1 we often use symbols rather than prose:

$$H_o : \mu = 100 \quad \text{versus} \quad H_1 : \mu \neq 100$$

$$H_o : \mu = 20 \quad \text{versus} \quad H_1 : \mu \leq 20$$

$$H_o : \mu = 40 \quad \text{versus} \quad H_1 : \mu \geq 40$$

Analogy to a Court Trial

- At the start of a trial, the defendant is assumed to be innocent, i.e H_o is assumed to be true.
- The alternative hypothesis is that the defendant is guilty: H_1.
- Evidence is provided to cast doubt on H_o (In a statistical test the evidence would be a random sample).
- In a trial a plaintiff (the party who initiates a lawsuit) attempts to use evidence to show that H_o is unlikely - "beyond reasonable doubt" - that is H_1 is more likely to be true.
- In statistics the standard for "beyond reasonable doubt" is the acceptable probability that the effect is not due to random variability in the data but is a real effect.
- In statistics we try to show that it was unlikely that the sample has the properties **by chance alone**.

Analogy to a Court Trial

- In statistics, it is never possible to prove that a null hypothesis is not true. This is true for most if not all scientific studies.

Analogy to a Court Trial

As with criminal trials it is possible to make mistakes.

- If a jury sends an innocent person to jail, in statistics that is called a **Type I error** (or false positive), i.e we claim that H_1 is true when it fact it is false.
- If a jury sets a guilty person free, in statistics this is called a **Type II error** (or false negative), i.e we claim that H_1 is false when in fact it is true.

	Innocent H_o is True	Guilty H_1 is True
Accept H_o - innocent	Correct	Type II Error (False **Negative**)
Reject H_o - guilty	Type I Error (False **Positive**)	Correct

Let's Put Some Data on Trial

- The container for shipping round worm medication weighs 48.0 grams
- 38 containers are picked at random and weighed. The average weight of the containers is 48.5 grams with a standard deviation of 1.2 grams.
- Determine whether the containers contain the claimed amount of 48.0 grams of medication at the 0.05 significance level
- We are going to try to answer the question, how likely is it to pick a sample where the weight is 48.75 gms
- Is it likely or unlikely?

State H_o and H_1

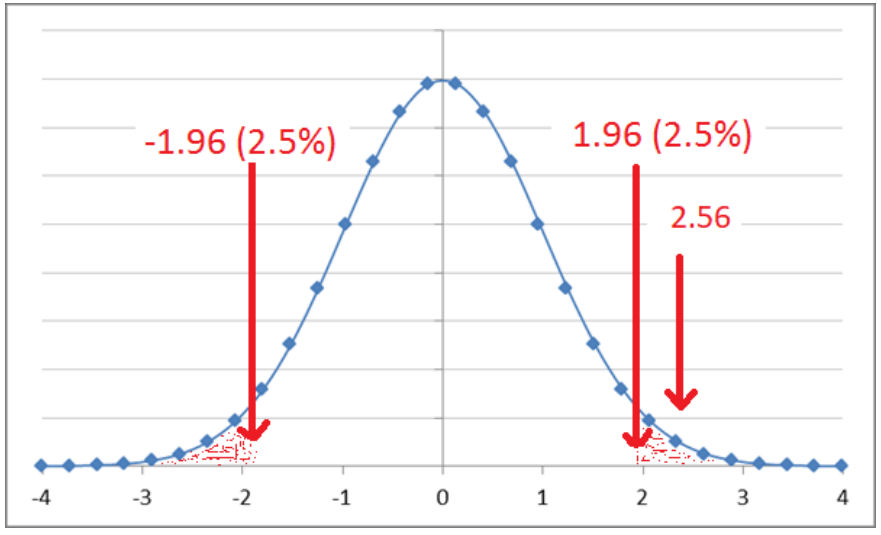
$$H_o : \mu = 48$$

$$H_1 : \mu \neq 48$$

Compute the standard error and Z statistic

$$\text{SE} = \frac{s}{\sqrt{n}} = \frac{1.2}{\sqrt{38}} = 0.195$$

$$z = \frac{48.5 - 48}{0.195} = 2.56$$



This is called a two-tailed test

Is the null hypothesis rejected?

- Since we are looking at the area bounded on either side, we must split the 0.05 significance level to be 0.025 for each tail.
- The area above 2.56 or below -2.56 is ~ 0.0052.
- The total area that includes both tails is $2 \times 0.0052 = 0.0104$. This value is less than 0.05, therefore the null hypothesis is rejected.
- It is likely that the containers were not packed correctly.

p-Values

- The area we found in the previous slide is called the **p-value**:

$$\text{p-value} = 0.0104$$

- The **smaller** the p-value the more likely we reject the null hypothesis.
- A low p-value is the probability of seeing the observation if it were simply a chance event.
- A low p-value suggests we more likely seeing a real effect.
- How small is small enough to reject H_o?

Ranges of p-Values

p-Value	Evidence against H_o
> 0.1	Very weak
Between 0.1 and 0.05	Weak
Between 0.05 and 0.01	Strong
< 0.01	Very Strong

Thresholds for p-Values

- When testing a hypothesis we normally set a threshold to determine significance.
- Ideally the threshold is set before one does the experiment.
- A common threshold is 0.95% and perhaps less often 0.99%
- If the p-value lies beyond the area determined by these, then H_o is rejected.

Mistakes

- As with any statistical test, a low p-value proves nothing, it just casts doubt on H_o.
- As we've seen, it is quite possible that a rejection of H_o is incorrect - **false positive**[2].

[2] We thought there was a real effect but there actually wasn't.

Let's Put Some Data on Trial

A pharmaceutical company buys raw material from Joes Cheap Chemicals in containers that are on average 1 Kg in weight.

The pharmaceutical company is suspicious that the containers it gets are consistently less than 1 Kg which allows Joes Cheap Chemicals to make more money by packing less material.

One Tailed Test

- The company measures 100 consecutive containers and finds that the mean weight is 0.95 Kg with standard deviation of 0.15 Kg
- Is this good evidence that the pharmaceutical company is being cheated?
- We are going to try to answer the question of how likely is it to pick **by chance** a sample where the weight of 0.95 is unusually low.
- Is it likely or unlikely?

Set up H_o and H_1

- The Null Hypothesis, H_o: Joes Cheap Chemicals is at the perfect weight

$$H_o : \mu = 1 \text{ Kg}$$

- The Alternative Hypothesis, H_1 : Joes Cheap Chemical containers are lighter than they should be:

$$H_1 : \mu < 1 \text{ Kg}$$

Single Tailed Test.

Compute Standard Error

- What is the probability that the average weight could by chance be less that 1 Kg?
- First compute the standard error of the sampled mean:

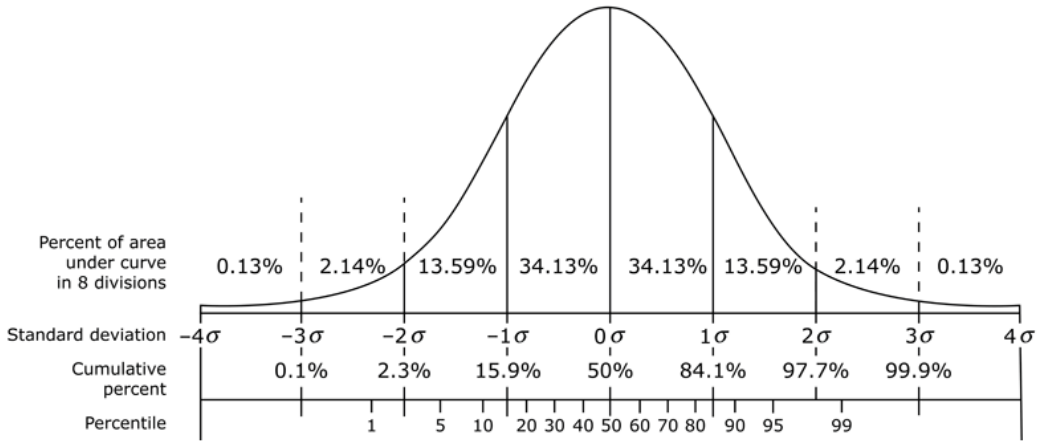
$$\text{SE} = \frac{s}{\sqrt{n}} = \frac{0.15}{\sqrt{100}} = 0.015$$

- 0.015 is the standard deviation of the means.
- What is the likelihood we could have obtained a mean of 0.95 from such a distribution?

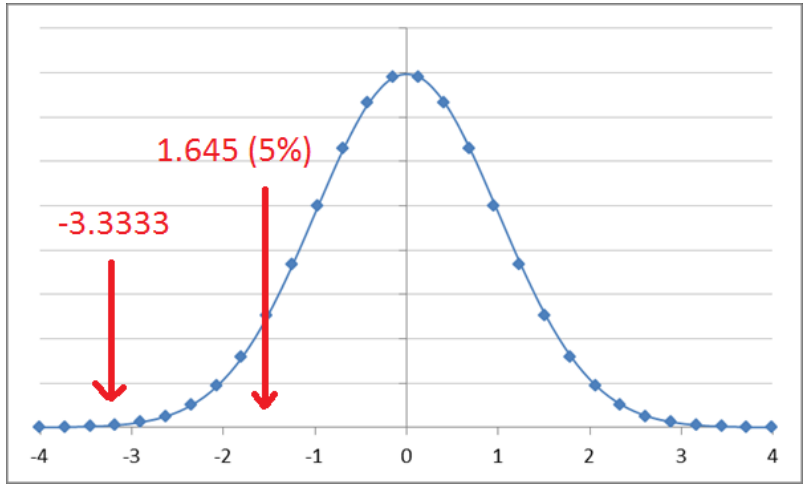
Standardize the Standard Error

- Standardize the value of the sample mean, $\overline{x}$, using the sample standard error of 0.015:

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{0.95 - 1}{0.015} = -3.333$$



Standardize the Standard Error



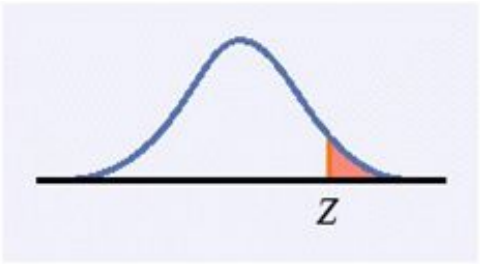
Find Area Below -3.333

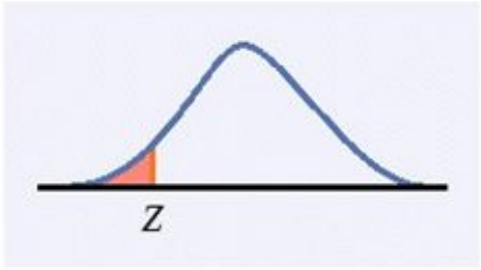
- Area below -3.333 $= 0.0005$

This means that there was only a 0.05% chance that this sample could have arisen **by chance alone**. This is highly unlikely. In other words getting a sample mean of 0.95 Kg is very unlikely by simple chance, it is much more likely that the supplier has tampered with the containers to reduce their weight.

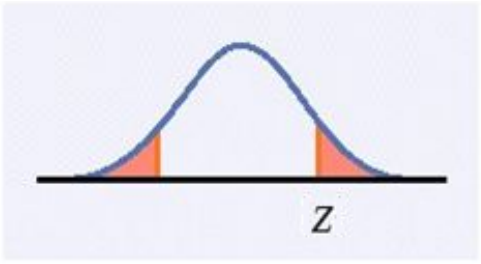
We therefore reject the null hypothesis that $\mu = 1$ Kg and instead accept the alternative hypothesis that $\mu < 1$ Kg.

Two-tailed and One-tailed Tests





One-tailed



Two-tailed

- In the first example (two tailed): $H_1 : \mu \neq 48$
- In the second example (one tailed): $H_1 : \mu < 1$ Kg

Example

- A population of cows was fed a special high-protein grain diet for a month.
- A random sample of 34 cows was weighed and it was found that the cows has gained on average 6.7 pounds with a standard deviation of 7.1.
- Test the hypothesis that the average weight gain per cow for the month was more than 5 pounds.
- State H_o, H_1, are we dealing with a single or two tailed test?

Example

We're testing the average weight gain:

$$H_o : \mu = 5$$

$$H_1 : \mu > 5$$

Example

$$z = \frac{6.7 - 5}{7.1/\sqrt{34}} = 1.396$$

What is the area beyond 1.396?

p-value $= 0.082$.

At a significance level of 0.05%, $0.082 > 0.05$, therefore picking 34 cows with an average weight increase of 6.7 could have happened by chance, therefore we do not reject H_o .

There is insufficient evidence to suggest that the high-protein diet increased the weight of the cows. The cows are innocent.

Example

A researcher claims that a viral drug delivery system manages to deliver into cells on average 1800 viruses per cell.

Forty cells are picked at random and the mean number of viruses detected per cell is found to be 1830 with a standard deviation of 200.

Does the evidence support the researchers claim?

State H_o, H_1, are we dealing with a single or two tailed test?

Example

$$H_o : \mu = 1800$$

''

$$H_1 : \mu \neq 1800$$

Use Two-tailed test

Example

Compute z-score:

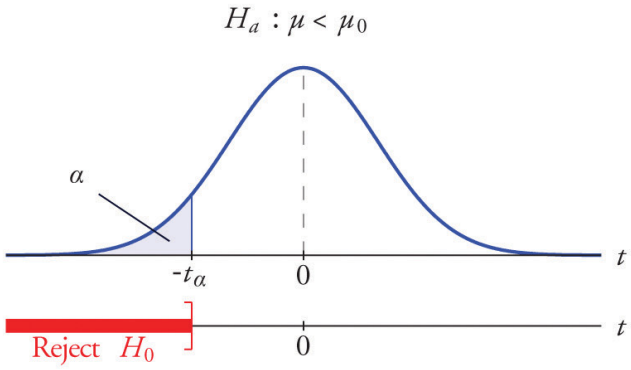
$$z = \frac{1830 - 1800}{233/\sqrt{40}} = 0.814$$

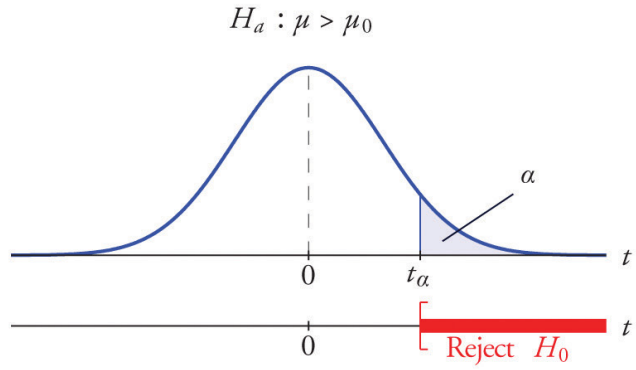
The area beyond this z value is 0.208. However because we need to do a two tailed test we must double this to include the same area in the left tail: Therefore:

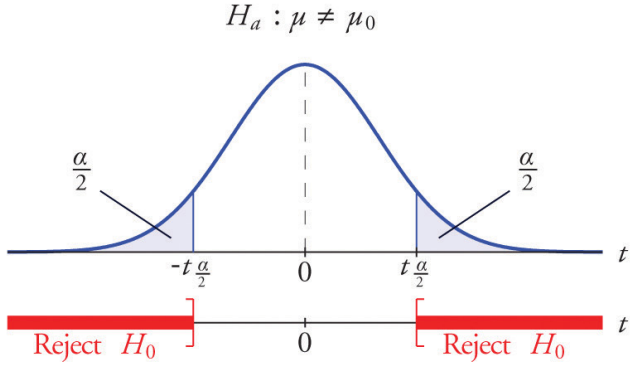
p-value $= 0.416$

This value is well within the acceptance area, therefore we no not reject H_o. We therefore believe the researcher.

Summary







`https://tinyurl.com/yb9fsmag`

Summary

Let $X_1, \ldots, X_n$ be a large ($n > 30$) sample from a population with mean μ and standard deviation σ.

To test the null hypothesis for the form:
$H_o : \mu \leq \mu_o, H_o : \mu \geq \mu_o,$ or $H_o : \mu = \mu_o$

- Set a critical value, α: eg 0.05, 0.01
- Compute the z-score, $z = \frac{\overline{X} - \mu_o}{\sigma/\sqrt{n}}$ If σ is unknown use sample standard deviation, s, instead.
- Compute the p-value (area under curve beyond the z score) according to:

Alternative Hypothesis	**p-value**
$H_1 : \mu \geq \mu_o$	Area to the right of z
$H_1 : \mu \leq \mu_o$	Area to the left of z
$H_1 : \mu = \mu_o$	Sum the area below $-z$ and above z

- If p-value is less than α then reject hypothesis, H_o.

Large Samples

You may have noticed that all the examples involved large samples.

Small Sample Tests for a Population Mean

If the sample size is less than 30 then we no longer test against a normal distribution but a t distribution. Instead of z we compute t:

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

with $n - 1$ degrees of freedom.

Example

The o-rings for an experimental heart value must have a diameter between 38.98 and 39.02 mm. The mean thickness of manufactured rings must be 39.00 mm. A sample of six rings is drawn and the diameter of each is measured. These diameters are 39.03, 38.997, 39.012, 39.008, 39.019 and 39.002.

With significance level of 95%, can we state that the sample most likely came from a population of rings with mean 39.00 mm?

Small Sample Tests for a Population Mean

$$H_o : \mu = 39.00 \qquad H_1 : \mu \neq 39.00$$

Because the sample size is small we must use a t distribution.

$$\overline{X} = 39.01133 \qquad s = 0.011928 \qquad n = 6 \qquad \text{df} = 6 - 1 = 5$$

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{39.01133 - 39}{0.011928/\sqrt{6}} = 2.327$$

This is a two-tailed test. What is the area under a t distribution beyond $t = 2.327$?

Small Sample Tests for a Population Mean

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

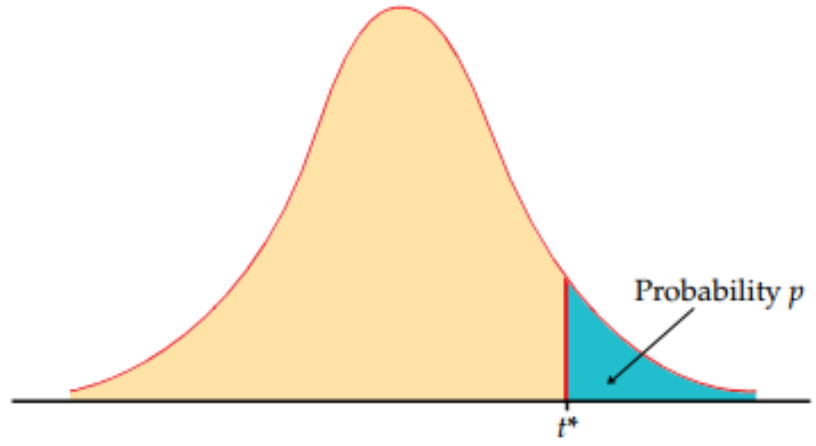


TABLE D

t distribution critical values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587

Small Sample Tests for a Population Mean

Use scipy's cdf function for the t distribution:

```
import scipy.stats as stats
print 1 - stats.t.cdf(2.327, 5)
0.03373
```

We double this value to get the total area for the two-tailed test.

p-value $= 0.0675$

This is just within the 0.05 critical value. On these grounds we don't reject H_o but since the result is on the edge, the manufacture is recommended to recalibrate their process.

Small Sample Tests for a Population Mean

If the population standard deviation, σ, is known then a z test can be done instead even if the sample is small. The problem with small samples is the poor estimate for σ and that's why we revert to a t-test in these circumstances.

Large Sample Tests for Difference between Two Population Means

Consider two samples from two independent populations. In each case we compute the mean. The question we wish to ask is are the means the same or different?

In other words, is $\mu_1 - \mu_2$ close to zero, or far from zero?

$$H_o : \mu_1 - \mu_2 = 0 \qquad H_1 : \mu_1 - \mu_2 \neq 0$$

From previous slides we know how the difference between two means is distributed:

$$X - Y \sim N(\mu_X - \mu_Y, \sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2) = N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y})$$

Large Sample Tests for Difference between Two Population Means

To test the null hypothesis we need to ask whether the difference between the means falls within or outside the critical region for the distribution:

$$X - Y \sim N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y})$$

Example

A experiment is conducted to investigate the amount of trace copper in the blood stream of human males and females.

A random sample of 75 males and 50 female donors yields concentration means of 28 and 33 ppm of copper respectively and standard deviations of 14.1 ppm and 9.5 ppm respectively. .

Test the hypothesis that the concentration of copper is the same in males and females.

Example

$$H_o : \mu_1 - \mu_2 = 0 \qquad H_1 : \mu_1 \neq \mu_2$$

This is a two-tailed test. Compute the z-score for the distribution of the difference in the means:

$$X - Y \sim N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y})$$

$$z = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_X^2/n_X + \sigma_Y^2/n_Y}} = \frac{33 - 28}{\sqrt{\frac{14.1^2}{75} + \frac{9.5^2}{50}}} = \frac{5}{\sqrt{2.65 + 1.81}} = 2.37$$

The area beyond 2.37 is 0.0089. Since it is a two tail test we double this area to get the total area in the two tails:

p-value $= 0.0178$

Since the p-value is less than 0.05 (i.e it is inside the critical area) we do not reject the null hypothesis.

There is no difference between females and males.

Small Sample Tests for Difference between Two Population Means

In many cases our samples are likely to be small. As a result we will often use the t distribution in our tests.

There are two cases to consider.

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$$

Small Sample Tests for Difference between Two Population Means

Case 1: $\sigma_X^2 = \sigma_Y^2 = \sigma^2$

Since we don't know the common variance, σ we must estimate it from the two sample variances by pooling the sample variances and weighting with respect to the sample sizes using:

$$s_p^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}$$

The t statistic is now computed from:

$$t = \frac{\overline{X} - \overline{Y}}{s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

with df $= n_X + n_Y - 2$

See example 6.14 in Navidi (p435)

Small Sample Tests for Difference between Two Population Means

Case 2: $\sigma_X^2 \neq \sigma_Y^2$

If the variances are unequal the degrees of freedom must be computed as follows:

$$\text{df} = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{(s_X^2/n_X)^2}{n_X - 1} + \frac{(s_Y^2/n_Y)^2}{n_Y - 1}}$$

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$$

See example 6.13 in Navidi (p433)