ST221 Introduction to Statistics

Dr Rafael Moral

Dept of Mathematics and Statistics
Maynooth University

rafael.deandrademoral@mu.ie

6 Statistical inference for continuous responses

Recap:

- Random variables: variables whose value depends on chance.
- Determine:
  - possible values X can take:
    e.g. $x = 0, 1, 2, ..., 10$, $x \in (0, 1)$, etc
  - probability mass function $p(x) = P(X = x)$, or probability density function $f(x)$.
- Note: so far we have assumed that the distribution and parameters are known. From now on, we use data to **estimate** the parameters of some assumed distribution. In this section we will examine continuous data and in the next section categorical data.

The general idea:

- We have some data (e.g. $X_1 = x_1, ..., X_n = x_n$) and want to ask some questions about the population the data came from.
- We assume the observations are random draws from some distribution where $E(X_i) = \mu$, $Var(X_i) = \sigma^2$.
- We use the data to estimate μ and σ^2.

Population versus sample.

- Terminology:
  - Population parameters: $E(X) = \mu$, $Var(X) = \sigma^2$.
  - Sample statistics:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

- Terminology:
  - Population parameters: $E(X) = \mu$, $Var(X) = \sigma^2$.
  - Sample statistics:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

- In the classical framework:
  - Parameters μ, σ^2 are some fixed, but unknown values that we would like to estimate from our data.
  - Statistics $\bar{X}$, S^2 are random variables.

6.1 Estimation and confidence intervals for large samples

Estimation

Example: a paper published in the Journal of the American Medical Association examined whether the true mean body temperature is 98.6 degrees Fahrenheit. The study gave 130 readings of body temperature for 65 women and 65 men.

The sample mean of these 130 readings was 98.249 and sample standard deviation 0.733. $\bar{x} = 98.249$, $s = 0.733$, $n = 130$.

Question: What is μ the population mean human body temperature?

Answer: Suppose the data represent a random sample from the population of all body temperatures. Then we would estimate the population mean μ by the sample mean $\bar{x}$, i.e. we **estimate** the mean human body temperature to be 98.249.

Reliability of the estimate of μ

*E*xample (human body temperature) contd.

We have estimated that the mean human body temperature is 98.249. But ***how reliable is this estimate***? If we repeated the study we would most likely get a different answer. But how different?

- We use the information about the variability of the data and the Central Limit Theorem (CLT) to answer that question.
- Suppose that the population has a standard deviation of σ. Our estimate of σ is the sample standard deviation s, which is 0.733.

Constructing the CI for μ

Estimate of μ: $\bar{x} = 98.249$, estimate of σ: $s = 0.733$.

- From the Central Limit Theorem:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \qquad \text{approximately for large } n.$$

- Using s^2 to estimate σ^2:

$$\bar{X} \underset{\text{approx}}{\sim} N(\mu, \frac{s^2}{n})$$

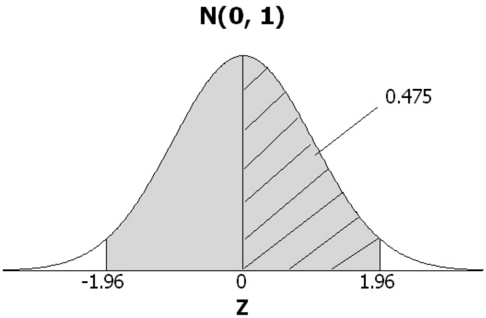
$$\bar{X} \underset{\text{approx}}{\sim} N(\mu, \frac{0.733^2}{130})$$

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \underset{\text{approx}}{\sim} N(0, 1)$$

- The quantity $\frac{s}{\sqrt{n}}$ is called the **standard error** of $\bar{X}$ and we write $SE(\bar{X})$.

- For the temperature data $SE(\bar{X}) = \frac{s}{\sqrt{n}} = \frac{0.733}{\sqrt{130}} = 0.064$.

- **FACT:** 95% of the area under the $N(0,1)$ curve lies in ± 1.96 (CHECK TABLES!). Or we can say $P(-1.96 < z < 1.96) = 0.95$



What does this mean?

$$P(-1.96 < z < 1.96) = 0.95$$

Therefore,

$$P(-1.96 < \frac{\bar{X} - \mu}{s/\sqrt{n}} < 1.96) = 0.95$$

And,

$$P(\bar{X} - 1.96\frac{s}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{s}{\sqrt{n}}) = 0.95$$

i.e. with probability 0.95 μ is contained in

$$\bar{X} \pm 1.96\frac{s}{\sqrt{n}}$$

which is called a **95% confidence interval** for μ.

Example contd. (Human body temperature)

$$\bar{x} \pm 1.96 \times \frac{s}{\sqrt{n}} = 98.249 \pm 1.96 \times 0.064 = (98.12, 98.38)$$

Interpretation: with 95% confidence, the true mean human body temperature is between 98.12 and 98.38.

Question: Is the true population mean really 98.6?

Answer: 98.6 is well outside of the 95% confidence interval, so it is very unlikely to be the true value of μ.

CI for large samples - general format.

In general, if we have a random sample of observations from a population we can estimate the population mean μ using $\bar{x}$ and if n is sufficiently large (usually $n \geq 30$) we can construct a $100(1-\alpha)\%$ confidence interval for the population mean μ using the formula

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

where $z_{\alpha/2}$ can be read from the tables.

- Will μ be in the confidence interval?
- Will $\bar{x}$ be in the confidence interval?

We say that we are $100(1-\alpha)\%$ confident that μ, the population mean will be in the interval.

Example.

Engineers from a motor company have developed a new type of break light for motor vehicles. As part of the safety evaluation program the company wished to estimate the mean drive response time to the new brake light. Fifty drivers were selected at random and the response time (in seconds) was recorded for each. The sample mean was 0.72 and the sample variance was 0.022. Estimate a 99% confidence interval for the mean driver response time to the new brake light.

Summary of estimation and confidence intervals for large samples

6.2 Estimation and confidence intervals for small samples

Estimation

Example: A pharmaceutical company has developed a drug they believe will help cure a particular disease and has acquired legal permission to test the drug on humans. However, there are safety concerns that the drug will have an adverse effect on blood pressure. They run an experiment to estimate the average increase in blood pressure for patients who take the new drug. They only have permission to test the drug on ten patients.

The ten recordings are 1.3, 2, 2.4, 1.1, 1.4, 3.2, 3.4, 2.4, 2.7, 3.1.
$\bar{x} = 2.3$, $s = 0.82865$, $n = 10$.

Example contd.

Question: What is μ the population mean increase in blood pressure for people who take the drug?

Answer: Suppose the data represent a random sample from the population of all people who take the drug. Then we would estimate the population mean μ by the sample mean $\bar{x}$, i.e. we **estimate** the mean increase in blood pressure to be 2.3.

Estimating the mean for a small sample is the same as for a large sample.

Reliability of the estimate of μ

When we have a large enough sample, we rely on the Central Limit Theorem to tell us that:

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \quad \underset{\text{approx}}{\sim} \quad N(0,1)$$

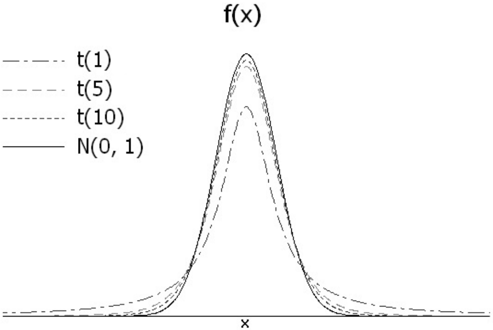
regardless of what distribution the population has.

If we have a small sample, we cannot make this assumption. However, in the case of small samples, if we know that the population the sample comes from is normal or approximately normal, then we can assume

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{(n-1)}$$

where $t_{(n-1)}$ denotes the t-distribution with $\nu = n-1$ degrees of freedom.

More about the t distribution



- t density curve is symmetric about 0 and bell shaped.
- only one parameter: degrees of freedom (ν).
- the density curve is flatter than N(0,1)
- ν controls how spread out the distribution is, i.e. how flat the density curve is.
- As ν increases, the t-density curve approaches the N(0,1) curve and the

Constructing the CI for μ

Therefore, if n is small we use the t-distribution to construct CIs:

$$P(-t_{(\nu,\frac{\alpha}{2})} < \frac{\bar{X}-\mu}{s/\sqrt{n}} < t_{(\nu,\frac{\alpha}{2})}) = 1-\alpha$$

so,

$$P(\bar{X} - t_{(\nu,\frac{\alpha}{2})}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{(\nu,\frac{\alpha}{2})}\frac{s}{\sqrt{n}}) = 1-\alpha$$

giving the interval,

$$\bar{X} \pm t_{(\nu,\frac{\alpha}{2})} SE(\bar{X})$$

where we read $t_{(\nu,\frac{\alpha}{2})}$ from the t-tables.

However, before constructing the CI, we should draw a histogram (or some other graphic) of the raw data to see if it follows a normal or approximately normal distribution. (This is not necessary if the sample is large enough because of the CLT.)

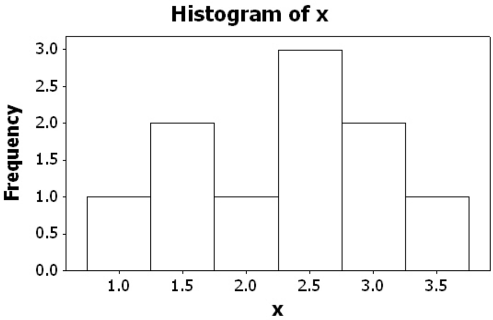
Note on the t-tables

- In standard normal tables, probabilities for a single density curve $N(0,1)$ are tabulated, while in Student's t tables critical values for 30 density curves t_ν, $\nu = 1, ..., 30$ are tabulated.

- In the t tables, the first column contains ν i.e. the degrees of freedom and the first row denotes the probabilities ($P(X > t_{(\nu, \frac{\alpha}{2})}) = \alpha/2$). The corresponding critical values ($t_{(\nu, \frac{\alpha}{2})}$) are in the body of the table.

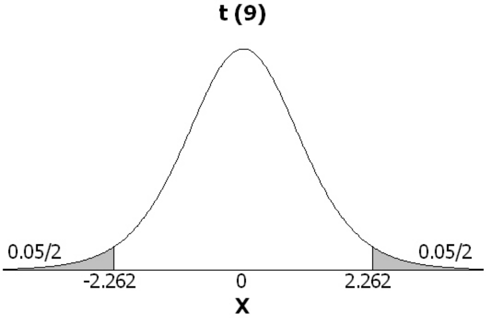
Example (pharmaceutical) contd.

- Do the data look roughly normal?



- Degrees of freedom $= \nu = n - 1 = 10 - 1 = 9$

- From the tables, 95% of the area under the t_9 curve lies between ±2.262.



Constructing the confidence interval:

The 95% confidence interval is given by

$$\bar{x} \pm t_{\nu, \frac{0.05}{2}} SE(\bar{x})$$

$$\Rightarrow 2.3 \pm 2.262 \frac{0.82865}{\sqrt{10}} = 2.3 \pm 2.262 \times 0.26204 = (1.707, 2.893)$$

Interpretation:

We are 95% confident that the population mean increase in blood pressure for people on the drug is between 1.707 and 2.893.

Additional thoughts:

Question: Is the true population mean equal to 0?

Answer: 0 is not in the 95% CI so it is not a plausible value for μ. The CI supports the claim that the drug has an adverse effect on blood pressure.

Example

Pulse rate is an important measure of the fitness of a person's cardiovascular system. The mean pulse rate for adult males is approximately 72 heart beats per minute. A random sample of 21 males who jog at least 10 kilometres per week had a mean pulse rate of 52.6 beats per minute and a standard deviation of 3.22 beats per minute. A histogram of the data was unimodal and symmetric. Construct a 90% confidence interval for the mean pulse rate of all adult males who jog at least 10 km per week.

Examples for small and large samples

Example

Data: 4, 7, 16, 12, 9, 8, 7, 6, 6, 11, 3, 19, 8, 7, 10, 7, 9, 6, 6, 13, 12, 16, 14, 7, 5, 5, 3, 8, 11, 15, 16, 11, 10.

Construct a 99% CI for μ.

Example

Data: 13, 19, 21, 24, 18, 17, 24, 19, 16, 22, 26, 19.

Construct a 95% CI for μ.

6.3 Hypothesis testing

The general idea:

- We have some hypothesis or theory about a population.
- We collect some data and use it to assess the truth of the hypothesis.

Example (blood pressure for people on a particular drug)

- Is μ the true mean change in blood pressure for people on the drug equal to 0?
  - Our null hypothesis is that the true mean change in blood pressure for people on the drug is equal to 0.
  - However, it is suspected that the true mean change in blood pressure is greater than 0. This is our alternative hypothesis.

- From these $n = 10$ observations we calculate the (sample) mean $\bar{x} = 2.3$ and the (sample) standard deviation $s = 0.82865$.

- $\bar{x} > 0$ but can we conclude that $\mu > 0$?

- Imagine that the population mean $\mu = 0$. Once we collect a sample and calculate its sample mean, we are unlikely to get exactly $\bar{x} = 0$.

Some definitions:

- The *null hypothesis* H_0 represents the statement being tested. In other words, this is a hypothesis we want to disprove.
- The *alternative hypothesis* H_A is what we suspect is true, i.e. what we want to establish.

Example (blood pressure) contd

- In this example, under our null hypothesis $\mu = \mu_0 = 0$ and we write:

  $$H_0 : \mu = 0 \qquad \text{versus} \qquad H_A : \mu > 0.$$

  Note: this alternative hypothesis is **one-sided**.

- If we suspected that the mean change in blood pressure μ is less than 0, we would write:

  $$H_0 : \mu = 0 \qquad \text{versus} \qquad H_A : \mu < 0.$$

  Note: this is also a **one-sided** alternative hypothesis.

- We could also have:

  $$H_0 : \mu = 0 \qquad \text{versus} \qquad H_A : \mu \neq 0.$$

  which is a **two-sided** hypothesis. In this case we have no prior belief for what μ should be, we just suspect it is not 0.

Deciding on the hypotheses

- We decide on our hypotheses **before** we see our data. The choice depends on the context. If you're unsure which hypothesis is appropriate, use the two sided alternative.

- In the context of testing a hypothesis about the population mean, the null hypothesis always has the form

$$H_0 : \mu = \mu_0$$

- The alternative hypothesis can take any of the following forms:

$$H_A : \mu \neq \mu_0$$
$$H_A : \mu > \mu_0$$
$$H_A : \mu < \mu_0$$

Test Statistic

A *test statistic* is used to compare the observed data with the null hypothesis H_0.

For testing hypotheses about μ the appropriate test statistic is:

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

If the null hypothesis is true (i.e. $\mu = \mu_0$ is correct):

- When $n > 30$, $T \sim \mathrm{N}(0, 1)$ approximately.
- When $n \leq 30$, $T \sim t_{n-1}$.
- In the case of $n \leq 30$, we assume that the population the sample came from is normally (or approximately normally) distributed.

Example (blood pressure) contd.

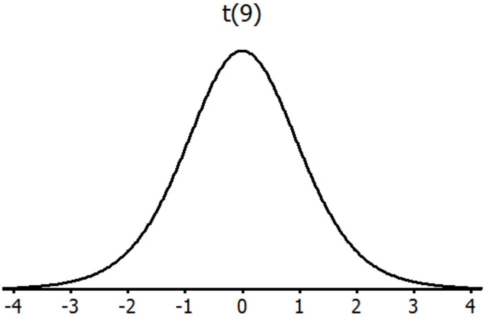
$H_0 : \mu = 0$ versus $H_A : \mu > 0.$ I.e. $\mu_0 = 0$

- The observed test statistic is:

$$T_{obs} = \frac{2.3 - 0}{0.82865/\sqrt{10}} = 8.78$$

- Does the data support the null hypothesis? Based on the observed test statistic we either:
  - Reject the null hypothesis and accept that the alternative hypothesis is true.
  - Or, fail to reject the null hypothesis and have no evidence that the alternative hypothesis is true.
- To decide which of these happens we compute the p-value.

Example (blood pressure) contd.



What is a P-value?

A p-value is the probability of observing the test statistic or a more extreme test statistic, given that the null hypothesis is true.

- Two-sided alternative: $H_0 : \mu = 0$ versus $H_A : \mu \neq 0$
  P-value $= 2\text{P}(t_{n-1} \geq |T_{obs}|)$.
- One-sided alternative: $H_0 : \mu = 0$ versus $H_A : \mu < 0$
  P-value $= \text{P}(t_{n-1} \leq T_{obs})$.
- One-sided alternative: $H_0 : \mu = 0$ versus $H_A : \mu > 0$
  P-value $= \text{P}(t_{n-1} \geq T_{obs})$.

If the P-value is very small, then our observed statistic is very extreme and H_0 is unlikely to be true. But what is 'very small'?

- Fix a *significance level* α.
- If the P-value is less than α then we reject H_0 at level α. Otherwise we fail to reject H_0.
- Use the significance level $\alpha = 0.05$ unless told otherwise.

Example (blood pressure) contd.

Here,

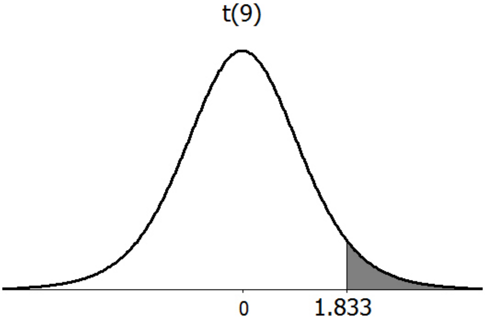
- One-sided alternative: $H_0 : \mu = 0$ versus $H_A : \mu > 0$
- We have observed $T_{obs} = 8.78$.
- P-value $=$ P$(t_{n-1} \geq T_{obs}) =$ P$(t_9 \geq 8.78)$.

We compute our p-value: P$(t_9 \geq 1.83) = 0.05$ and since $8.78 > 1.83$, we know that P$(t_9 \geq 8.78) < 0.05$.

Conclusion: We reject the null hypothesis and conclude that $\mu > 0$.

NB: In the case that we fail to reject the null hypothesis that does not mean that we have proven that it is true. Just that we have no evidence for the alternative. We therefore do **not** say that we accept the null hypothesis.

Example (blood pressure) contd.



Additional examples of computing a p-value

Example

$H_0 : \mu = \mu_0$, $H_A : \mu > \mu_0$,

$n = 15$

$T_{obs} = 0.75$

P-value

Example

$H_0 : \mu = \mu_0,\ H_A : \mu < \mu_0,$

$n = 54$

$T_{obs} =$ -2.73

P-value

Example

$H_0 : \mu = \mu_0,\ H_A : \mu \neq \mu_0,$

$n = 18$

$T_{obs} = 1.5$

P-value

Steps in a hypothesis test

If the sample size is below 30, we first need to be sure that the sample comes from an approximately normally distributed population. If the sample size is large enough, the CLT tells us about the sampling distribution of the mean, regardless of the underlying distribution of the population. The steps are:

1. Set up the null and alternative hypotheses
2. Compute the test statistic
3. Compute the P-value
4. State conclusions

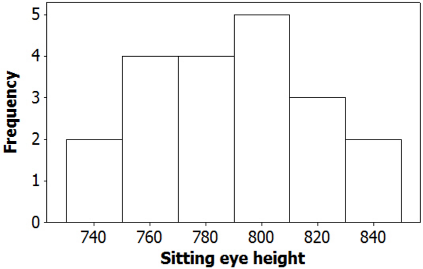
Example

A lab is asked to analyze some vitamin tablets to check if the active ingredient has a mean concentration of 0.86. The lab tests 35 tablets. They find that the mean concentration is 0.8404 with a standard deviation of 0.05. What can you conclude from this?

Let μ be the mean concentration in the population of tablets.

Example

When designing a cinema, engineers consider the 'sitting eye' heights of the audience. A random sample of the eye sitting height (in mm) of 20 men at a cinema was recorded. The summary statistics are $\bar{x} = 787.95$ and $s = 30.35$.



Is there evidence that the mean sitting eye height is above 750?

6.4 Comparing two independent samples

- So far, we have looked at inference for a population mean based on a sample.
- Suppose we have two samples from two different populations and we want to compare them.
  - Data $x_1, x_2, ..., x_n$ are random sample from the first population and data $y_1, y_2, ..., y_m$ are a random sample from the second population.
  - The samples are independent.

Example

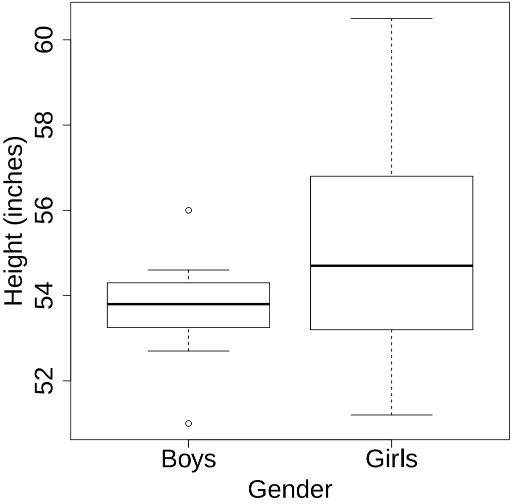
The heights of a group of children all age 10 were recorded.

- Assume these represent independent random samples from populations of all 10 year old girls and boys.
- Let the first population have mean μ_X and variance σ^2.
- Let the second population have mean μ_Y and variance σ^2.

The data from the experiment are:
Heights of girls: 51.2, 51.7, 53.2, 54.2, 54.3, 55.1, 56.7, 56.8, 57.2, 60.5
Heights of boys: 51.0, 52.7, 53.8, 53.8, 54.0, 54.6, 56.0.

- We get the summary statistics:
  - Girls: $\bar{x} = 55.09$, $s_X = 2.80573$ $n = 10$
  - Boys: $\bar{y} = 53.7$, $s_Y = 1.55456$, $m = 7$.

Example Heights data contd.



Estimation

- How do we estimate the difference between μ_X and μ_Y the means for each population?
  - We estimate μ_X by the sample mean $\bar{x}$ and we estimate μ_Y by the sample mean $\bar{y}$.
  - It follows that $\bar{x} - \bar{y}$ is an estimate for μ_X - μ_Y.
- How do we estimate the variance σ^2?
  - Remember that we have assumed that both populations have the same variance σ^2.
  - We can estimate this from either s_X and s_Y.
  - However, it would be better to use both to estimate σ^2.
  - We 'pool' these two statistics into an overall estimate of σ^2:

$$s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}.$$

Confidence interval for the difference between two means

A confidence interval for $\mu_X - \mu_Y$ is given by

$$(\bar{x} - \bar{y}) \pm t_{(\nu, \frac{\alpha}{2})} \times SE$$

- The SE is the standard error (remember for the one-sample CI this was $s/\sqrt{n}$). Here the SE$(\bar{X} - \bar{Y})$ is

  $$s_p\sqrt{\frac{1}{n} + \frac{1}{m}}$$

  where s_p is the pooled estimate of σ.

- $t_{(\nu, \frac{\alpha}{2})}$ depends of the confidence level and the degrees of freedom which are given by df$=n + m - 2$.

Example (heights of 10 year olds) contd.

Back to our data, we had:

- Girls: $\bar{x} = 55.09$, $s_X = 2.80573$ $n = 10$
- Boys: $\bar{y} = 53.7$, $s_Y = 1.55456$, $m = 7$.
- $s_p^2 = 5.6899$ and $s_p = 2.3854$, df = 10+7-2=15.

To construct a 95% CI for $\mu_X - \mu_Y$:

$$(\bar{x} - \bar{y}) \pm t_{(\nu, \frac{\alpha}{2})} \times SE = (-1.1, 3.9).$$

Interpretation: We are 95% confident that in the population of 10 year olds, the average height for girls is between 1.1 inches shorter to 3.9 inches taller than the average height for boys.

Example (heights of 10 year olds) contd.

In this example, we have made the following assumptions:

- Both populations have the same variance
- Both populations are normal.

Hypothesis test to compare two populations

The steps are the same as before:

1. Set up the null and alternative hypotheses
2. Compute the test statistic
3. Compute the P-value
4. State conclusions

In general in hypothesis testing, the test statistic is of the form:

$$T_{obs} = \frac{\text{estimate - parameter value under null hypothesis}}{\text{SE}}$$

Example (heights of 10 year olds) contd.

Following the steps:

1. If our goal is to establish that ten year old girls are taller, we write: $H_0 : \mu_X - \mu_Y = 0$ vs $H_A : \mu_X - \mu_Y > 0$.

2. The observed test statistic is:

$$T_{obs} = \frac{\bar{x} - \bar{y} - (0)}{SE} = 1.39/1.1755 = 1.18$$

3. P-value $= P(t_{15} \geq 1.18) > 0.05$.
   Since: $\text{P}(t_{15} \geq 1.753) = 0.05$.

4. Since P-value > 0.05 we fail to reject H_0. There is no evidence that the population mean heights of 10 year olds girls is larger than for boys.

Example

An experiment compared the effects of two growth hormones on the weight gain of pregnant rats. Weight gains are recorded for 20 rats receiving hormone A and 20 receiving hormone B.

The data are summarized as:
A: $\bar{x} = 51.8$, $s_x = 10.6$,
B: $\bar{y} = 60.2$, $s_y = 16.4$.

Test the appropriate hypothesis and give the 95% CI for the mean difference in weight gain between the treatment populations.

The hypothesis test:

Example contd.

The confidence interval:

6.5 Comparing two dependent samples

Two samples with paired data

- Up to now we have assumed that the two samples are independent of each other.
- What if this is not the case?

Example Does medication M reduce the blood pressure of the user?

- An experiment was carried out to test this. The blood pressures of 15 women were recorded prior to starting the treatment. The data is: $x_1, x_2, ..., x_{15}$.
- After taking the medication M for 6 months their blood pressure was recorded again. The new data is: $y_1, y_2, ..., y_{15}$.
- Here the x and y samples are not independent, because we have two measurements on each person.
- We cannot use the methods for two independent samples.
- Method: calculate differences $d_i = x_i - y_i$. Thus our new data is: $d_1, d_2, ..., d_{15}$. Analyse these using one sample methods.

Example

Do self reported (SR) and measured (M) heights differ?

A study of 12 males aged 12-16 asked them for a self assessment of their height and also measured the height in inches. The results are:

SR	68	71	63	70	71	60	65	64	54	63	66	72
M	67.9	69.9	64.9	68.3	70.3	60.6	65.5	67.0	55.6	74.2	65.0	70.8

6.6 Statistical inference for categorical responses

In the previous sections we looked at tests for continuous data for single variables (e.g. the weight of a person) and also compared continuous data for two groups (e.g. comparison of weights for males and females).

In this section we will examine categorical variables with 2 categories or levels.

Examples of categorical data (where k is the number of categories or levels):

- Gender: male or female (k=2)
- Drink alcohol: a lot, a little, not at all (k=3)
- Cause of death: heart disease, cancer, accident, other (k=4).

When the number of levels is two, we call this a binary variable.

A single binary variable

Example

500 people chosen at random were surveyed and 315 supported a government policy on a particular austerity measure. Here, each person answers either yes or no, so k=2. Let X be the number of people in the sample who answered yes.

- Prior to collecting the data, we know that $X \sim \text{Binomial}(n, p)$, where n = sample size and p = probability that a person says yes.
- p is the proportion of the population that supports the policy.

Generally we don't know p but it is something that we would like to know and this is often the motivation for carrying out the survey in the first place.

Estimation

We can estimate p from the data:

$$\hat{p} = \frac{x}{n}$$

In our example, $\hat{p} = \frac{315}{500} = 0.63$.

So we estimate that 63% of the population supports the government policy. But how reliable is this estimate?

NB: What is the difference between p and $\hat{p}$?

Reliability of the estimate of p

We can construct a confidence interval for p. Recall: $X \sim \text{Binomial}(n, p) \approx \text{N}(np, np(1-p))$ for large n. Therefore,

$$X \underset{approx}{\sim} N(np, np(1-p))$$

$$\text{So,} \quad \frac{X - np}{\sqrt{np(1-p)}} \underset{approx}{\sim} N(0,1)$$

$$\text{So,} \quad \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \underset{approx}{\sim} N(0,1)$$

Since we don't know p, we use

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \text{and} \quad \frac{\hat{p} - p}{SE(\hat{p})} \underset{approx}{\sim} N(0,1) \text{ for large } n.$$

Confidence interval for p

When n is large, we get a 100(1-α)% CI for p using:

$$\hat{p} \quad \pm \quad z_{\frac{\alpha}{2}} SE(\hat{p})$$

I.e.,

$$\hat{p} \quad \pm \quad z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Assumptions for the interval to be valid:

- Need the sample to be a simple random sample from a large population.
- Need both the expected number of ‘successes’ and ‘failures’ to be at least 10.

Example contd. (government policy)

Interpretation

Hypothesis testing for p

In general we have:

$$\text{H}_0\text{: } p = p_0$$

versus

$$\text{H}_A\text{: } p \neq p_0$$

or

$$\text{H}_A\text{: } p > p_0$$

or

$$\text{H}_A\text{: } p < p_0$$

For the confidence interval, we computed $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, however, for the hypothesis test we construct our test statistic assuming that the H_0 is true and so we use

$$\sqrt{\frac{p_0(1-p_0)}{n}}$$

as the denominator of our test. To test a population proportion, the test statistic therefore is:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \underset{approx}{\sim} \mathrm{N}(0,1) \;\text{ under } \mathrm{H}_0$$

Assumptions (as before):

- Need the sample to be a simple random sample from a large population.
- Need both the expected number of ‘successes’ and ‘failures’ to be at least 10.

Example contd. (government policy)
The government want to establish whether more than 55% of the population support their policy. We can do a formal hypothesis test for this.