11. - 12. Uncertainty and statistical models

Contents

Introduction

These notes consider Chapters 11 (statistical models and dealing with uncertainty) and 12 (statistical tests and hypothesis testing) of the handbook. These chapters are more theoretical with little code examples so I decided to combine them. There are, however, some formulae and concepts worth writing down.

Sample mean as a simple model

See page 227.

Suppose we have n measurements $x_1, x_2, \ldots, x_n$ with a sample mean of $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The simplest estimate for the error is the Sum of Squares (SoS) defined as

$$SoS = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Squaring makes sure that each term of the sum stays non-negative and, therefore, describes the squared "distance" of each measurement to the sample mean. The SoS is a very rough estimate and not that usable as the value tends to increase while n increases. A better estimate is the Mean Squared Error (MSE), which is the SoS divided by n. In other words, MSE is the mean of the squares of the errors defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

This estimate is already better but in the wrong dimension (squared) compared to the measurements. More meaningful estimate of the error can be achieved by taking the square root of the MSE i.e. $\sqrt{MSE}$.

Looking back to Chapter 6 (pages 116-117) we can see that the MSE is the population variance (σ^2) and $\sqrt{MSE}$ is the standard deviation (σ) **if** the measurements cover the whole population.

Law of large numbers and central limit theorem

See page 230. Formal definitions mentioned here are partially supplemented from Wikipedia.

Suppose we have an infinite sequence of independent and identically distributed random variables $X_1, X_2, \ldots, X_n$ with expected values $E(X_1) = E(X_2) = \cdots = \mu < \infty$. The law of large numbers (LLS) states that the sample mean

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

converges to the expected value of the population i.e.

$$\bar{X}_n \to \mu \text{ as } n \to \infty.$$

In other words, the expected value of the population can be estimated using the sample mean if the amount of samples is large enough. This, however, does not tell us anything about the precision of the estimate. In practice, it is impossible to draw infinite amount samples meaning some uncertainty always remains.

In addition, the central limit theorem (CLT) states that, as $n \to \infty$

$$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n}),$$

where σ^2 is the variance of the population.

In summary, as $n \to \infty$, the LLS states that the sample average converges to the expected value of the population and the CLT states that the distribution of $\bar{X}_n$ gets arbitrarily close to the normal distribution *regardless*[1] of the original distribution of X_i. Additionally, we now have an error estimate for the precision of the mean for the finite sample size, namely the standard error of the (sample) mean (SEM)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

which is the standard deviation of the distribution of $\bar{X}_n$. The SEM is the estimate for how close the sample mean is to the expected value of the population. We can immediately see that the precision increases when the sample size increases or the variance of the population decreases. Note, though, that the the increase in precision by drawing more samples is not linear but slower i.e. a square root of n.

Confidence intervals

See page 234.

By definition of the normal distribution, about 68.27 % of the values are within one standard deviation from the expected value ($\mu \pm \sigma$). This is illustrated in the figure below. A confidence interval (CI) is the interval where a parameter being estimated can typically be found (with a certain confidence level). For instance, in case of our expected value, 68.27 % of the possible values for the sample mean are within one SEM from the expected value of the population. This **does not** mean that the probability to find the expected value of the population within $\pm\sigma$ interval is 68.27 %! It merely states that, if an infinitely many samples were drawn, a sample mean would be within the CI 68.27 % of the cases. Of course, more sensible confidence levels are usually used. For instance, a 95 % confidence level corresponds to roughly $\pm 2\sigma$ interval (1.96 to be more precise). Additionally, when the standard deviation of the population is not known (required to calculate the SEM), the Student's t-distribution can be used to model the distribution of the sample means.

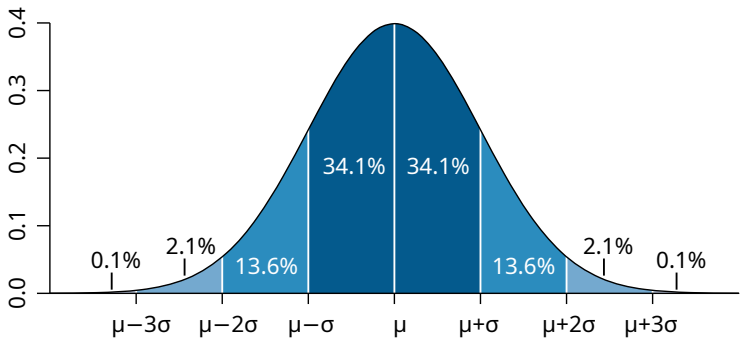


Figure 1: For the normal distribution, the values less than one standard deviation away from the mean account for 68.27 % of the set; while two standard deviations from the mean account for 95.45 %; and three standard deviations account for 99.73 %. (Wikipedia/Ainali)

[1] See Figure 11.6. in page 231.

Hypothesis testing

See page 243 onwards. Some additional theory is from Wikipedia.

In order to test a scientific hypothesis it is important to formulate a more concrete and measurable (i.e. numerical) statistical hypothesis that can be used to test hypotheses on the population level using smaller samples. This is done by setting up a null hypothesis H_0 and an alternative hypothesis H_1. They should be mutually exclusive. For instance: "A new medicine lowers the blood pressure (BP) on patients." is an understandable statement but cannot really be validated as-is.

The validity of this statement could be tested with the following experiment. A large number of people are selected. Half of them will receive the new medicine (experiment group) and half of them a placebo (control group). Blood pressures are measured initially and after the medicine should have taken an effect ($\Delta\bar{BP} = \bar{BP}_{end} - \bar{BP}_{initial}$). The statistical hypotheses would then be:

- H_0 : The new medicine does not lower the blood pressure more than the placebo in population i.e. $\Delta\bar{BP}_{medicine} = \Delta\bar{BP}_{placebo}$
- H_1 : The new medicine does lower the blood pressure more than the placebo in population i.e. $\Delta\bar{BP}_{medicine} < \Delta\bar{BP}_{placebo}$

Now, what is the point of all of this? Setting statistical hypotheses up this way many biases, such as placebo effects and the Texas sharpshooter fallacy (i.e. coming up with hypotheses *after* seeing the data), can be mitigated. Note that these tests can merely estimate which hypothesis is more supported by the dataset.

The null hypothesis does not indicate that something is "wrong" and the alternative is "right". If H_0 is accepted after testing, it just means that the effect being studied (assumed difference in two variables) does not exist in population or it is observed purely by change in a particular sample. H_0 could even be the scientific hypothesis to be tested e.g. "A new medicine should not decrease the blood pressure as a side effect".

Statistical significance of a test can be expressed with a p-value ($p \in [0, 1]$) which describes the probability that the observed result is obtained assuming H_0 is correct. In other words, a small p-value indicates that the observation is unlikely under H_0. In order to calculate any estimates, however, the null distribution (i.e. the distribution assuming H_0 is true) must be known. Additionally, there are some pitfalls when interpreting p-values:

- The p-value is *not* the probability that H_0 is true (of H_1 false). It is the probability that, under H_0, the observed sample is acquired.
- The 0.05 significance level (which is oftentimes used to reject H_0 if $p < 0.05$) is a convention! Consider using a lower value of 0.01, 0.005, or 0.001.
- Rejection of H_0 does not mean it is *false*. Rejection merely states that the data, assuming a certain distribution to model it, does not support this hypothesis. On top of the p-value, rejection or acceptance should *ideally* be based on additional criteria (e.g. data quality) as well.

Due to the pitfalls mentioned above or e.g. sampling or measurement errors, a wrong conclusion can happen. Two types of errors can happen when using hypothesis testing:

- **Type 1 error** happens when an alternative hypothesis is incorrectly accepted
- **Type 2 error** happens when the null hypothesis is incorrectly accepted.

Type 1 error can be considered to be worse of the two due to the publication bias - a new result is generally easier to get published and could, therefore, steer a research topic to a wrong direction for a long time. Type 2 error, although bad in itself, is less dangerous as there might be some other studies in the future that come to the correct conclusion. Type 1 error can be mitigated by lowering the p-value limit for accepting the alternative hypothesis. This, of course, increases the chance for the type 2 error.

Permutation testing

If not already done, install the example datasets (should be installed automatically with the R environment, though).

```
install.packages("remotes")
library(remotes)

install_url(
  "http://emotion.utu.fi/wp-content/uploads/2019/11/nummenmaa_1.0.tar.gz",
  dependencies=TRUE
)
```

Dataset *vilja* (grain) is needed here. First, let's make sure it works.

```
library(nummenmaa)
vilja
```

```
##         RYHMA SATO
## 1         Koe   12
## 2         Koe   11
## 3         Koe   10
## 4         Koe    9
## 5         Koe    8
## 6  Kontrolli    7
## 7  Kontrolli    8
## 8  Kontrolli    9
## 9  Kontrolli   10
## 10 Kontrolli    6
```

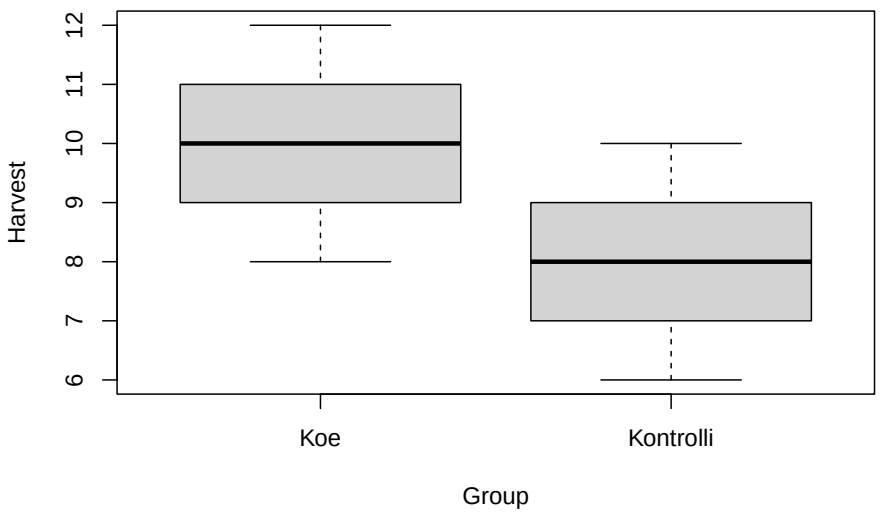
It should contain two columns *RYHMA* (group) and *SATO* (harvest). Former has two options *Koe* (Experiment) and *Kontrolli* (Control), latter is an integer value indicating the amount of harvest in kilograms.

There are ten equally large pieces of farmland. On half of them (selected randomly), a new fertilizer is used. The rest are left as a control group to estimate whether or not the fertilizer works. Let's formulate the statistical hypotheses.

- H_0 : fertilizer does not affect the harvest ($\mu_{experiment} = \mu_{control}$)
- H_1 : fertilizer increases the yield of the harvest ($\mu_{experiment} > \mu_{control}$)

Let's explore the data.

```
# Boxplot harvest by group
boxplot(vilja$SATO~vilja$RYHMA, xlab = "Group", ylab = "Harvest")
```



By visual inspection only, the overall yield of the experiment does indeed appear to be larger than in the control group. But is this result statistically significant? After all, there is noticeable overlap in the ranges of the two groups.

The idea of the permutation testing is to shuffle the observed data in order to generate a null distribution. This can be then used to estimate how likely it would be to get the observed data assuming the null hypothesis. The first step is to define the test quantity - this can be almost anything that can be calculated from the sample. In our case, we are interested in the difference in the means of the two groups, namely

$$\Delta\mu = \mu_{experiment} - \mu_{control}.$$

The value for $\Delta\mu$ is 2 in the original data.

```
tapply(vilja$SATO, vilja$RYHMA, summary)
```

```
## $Koe
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       8       9      10      10      11      12
##
## $Kontrolli
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       6       7       8       8       9      10
```

Now, the permutations are done by keeping the yields as they are but shuffling the groups and calculating $\Delta\mu$s for each permutation. In other words, it is now randomized whether or not a yield measurement belongs to the experiment or control group. With enough iterations, this forms the null distribution. In case the H_0 is true, it should not matter where the seeds were planted as the fertilizer would not make difference and the observed $\Delta\mu$ would be a typical value in the null distribution. On the other hand, if it would be unlikely to get the observed $\Delta\mu$ under H_0, then the fertilizer probably made a difference.

Note that making permutations is tedious. Even with ten rows a total of $10! = 3628800$ permutations exist.

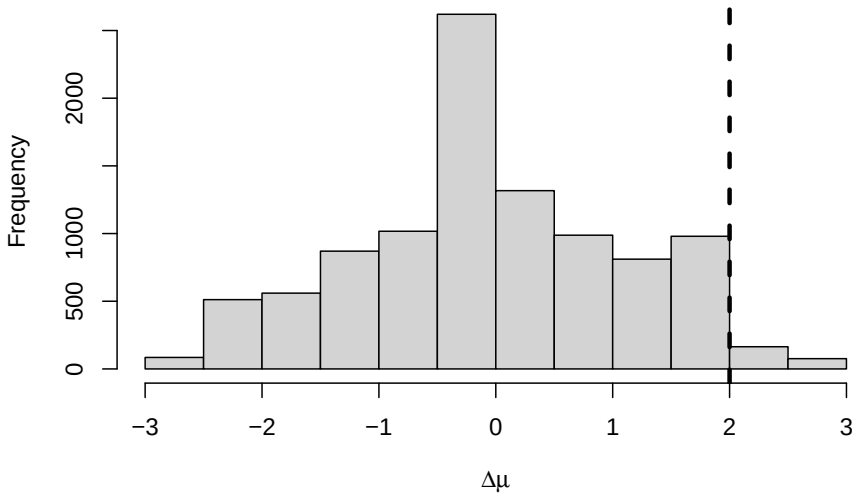
```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```r
# Observed difference in the means (2)
observed <- mean(vilja[vilja$RYHMA == "Koe", "SATO"]) -
  mean(vilja[vilja$RYHMA == "Kontrolli", "SATO"])

# Set the same seed as in the example (p. 248) to verify
set.seed(271142)

# Number of permutations
n_iterations <- 10000

# Initialize the results vector
results <- numeric(n_iterations)

# Run permutations
for (i in 1:n_iterations) {
  # Shuffle groups
  new_rows <- sample(nrow(vilja))
  new_data <- transform(vilja, SATO=SATO[new_rows])

  # Print the first permutation as an example
  if(i == 1){
    print(new_rows)
    print("Former row 1 (value 12) is now in row 3 etc.")
    print(new_data)
  }

  # Calculate difference for the permutation and save the result
  results[i] <- mean(new_data[new_data$RYHMA == "Koe", "SATO"]) -
  mean(new_data[new_data$RYHMA == "Kontrolli", "SATO"])
}
```

```
##  [1]  4  7  1  8  3 10  9  5  6  2
## [1] "Former row 1 (value 12) is now in row 3 etc."
##        RYHMA SATO
## 1        Koe    9
## 2        Koe    8
## 3        Koe   12
## 4        Koe    9
## 5        Koe   10
## 6  Kontrolli    6
## 7  Kontrolli   10
## 8  Kontrolli    8
## 9  Kontrolli    7
## 10 Kontrolli   11
```

```r
# Plot the results
hist(results, xlab = expression(paste(Delta, mu)), main = NULL)
abline(v = observed, lty = 2, lwd = 3) # Dashed line at the observed value
```



As should be, the distribution is centred at 0 (which is the case that there is no difference in the means, i.e. H_0). However, already by eye we can see that value 2 is close to the edge of the distribution implying that it is unlikely to get the observed value under H_0. But is the result statistically significant?

As o one-sided test, we can calculate how likely it would be to get a value of 2 or higher from the null distribution by chance (divide the amount values >2 with the amount of permutations). This states that 2.4 % of the values in the null distribution are larger than 2 indicating that it is unlikely to get the observation by change so the new fertilizer probably had an effect.

```
sum(results > observed) / n_iterations
```

```
## [1] 0.024
```

Like practically everything, the permutation testing method has pros and cons. The good thing is that it is quite versatile and can be used in many cases without the need to know the null distribution etc. beforehand. The biggest problem with this method is, however, the need for computational power - especially with larger datasets. Even with this small example and relatively low amount of iterations the calculations are not instant.

Testing with distribution models

See page 250.

Faster and more convenient way to setup tests is to use well-defined statistical distributions to model the null distribution. This requires stricter assumptions to be made. First of all, the distribution for the tested quantity needs to be known. Oftentimes this is the case so it is possible to immediately proceed to estimate the probability that an observed result is acquired assuming H_0 is valid. Most common statistical methods can be divided into three main categories:

- **Comparison of sample means** can be used to estimate whether or not some measurable quantity has different values between different samples. Examples: ***t*-tests** and **variance analysis** methods.

- **Covariance studies** are used to estimate whether or not the variability of two or more variables is joined. Examples: **correlation coefficients** and **regression analysis**.
- **Classification and dimension reduction** methods are useful when there are dozens or hundreds of variables to be estimated. These methods can reveal dependencies between variables in the data but are not suitable for hypothesis testing. Example reduction methods: **principal component analysis** and **factor analysis of mixed data**. Example classification methods: **logistic regression analysis** and **cluster analysis**.

These types of tests are known as parametric tests i.e. tests that make assumptions about the parameters of the underlying distribution. In case the assumptions (usually normal distribution and precise enough scale) cannot be fulfilled, there are also non-parametric tests to handle these cases. They are more robust but have less statistical power i.e. larger sample sizes are needed. Additionally, many non-parametric tests are based on the order, not values of the data (e.g. largest value is the first etc.).

Examples

A few simple test methods are explained below to give a general idea about parametric tests. See page 254.

Confidence interval of the mean

Assume we know the expected value and, say, the 95 % confidence interval of some (normal) distribution on the population level. Now, if a measured average of some sample is not within this interval ("within the error bars"), the measurement would be statistically significantly ($p < 0.05$) larger or smaller than the expected value.

Other example: assume we have two different measurements (means) from the same population with known 95 % confidence intervals. In case the intervals do not overlap ("overlapping error bars"), it can be said that the distribution of these two means overlap *at most* by 5 %.

Note, though, that the opposite in these cases does not hold. For instance, overlapping intervals do not imply that the means are equal (they *can* be, of course).

Z-tests

Standard normal distribution is assumed throughout this section. See page 255.

The visual inspection methods described above can be formulated mathematically using Z-tests. The one-sample Z-test is used in the former scenario: an observed mean of a sample is compared to the expected value of the population. The Z-score is defined as

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}.$$

This describes the difference between the sample and population ($H_0 : \bar{x} = \mu$) averages normalized by the SEM meaning the Z-score follows the standard normal distribution and directly gives the difference in the amount of standard deviations.

For example: the measured capasity of the working memory in a population is described by $\mu = 5$ and $\sigma = 2$ in some arbitrary units. Now, a sample of $n = 50$ people are randomly selected and a mean of $\bar{x} = 5.7$ is observed for the capacity. Using the equation above, we get $z = 2.5$ meaning the sample average is 2.5σ above the μ. By looking at Fig. 1, we can see that the observed value lands at the end tail of the distribution implying that the result probably does not support the null hypothesis.

In order to get a numeric result, we can use the PDF to calculate what this means exactly. Usually, and now, a two-sided test is used meaning half of the rejection area is in the lower and half in the upper tail of the distribution. In other words, if 95 % confidence level is used, 2.5 % of the lowest and 2.5 % of the highest values are used for rejection. If a one-sided test was used, the whole 5 % of the values in the lower OR upper tail of the distribution would be used as a rejection area. One-sided test is more powerful but should only be used when one is absolutely sure that the observed quantity should be larger or smaller than the expected value ($> \mu$ or $< \mu$). The two-sided test "goes both ways" ($\neq \mu$) so it tests both directions simultaneously and is not "blind" to the chance that a wrong direction was assumed.

So, how unlikely would it be to get this result? Let's calculate the PDF using the Z-score

```
# CDF, upper tail
pnorm(2.5)
```

```
## [1] 0.9937903
```

This means that 99.38 % of the values in the distribution are lower than 5.7 and, therefore, 0.62 % equal or higher. This can also be calculated using the CDF and symmetry of the distribution

```
# CDF, lower tail
pnorm(-2.5)
```

```
## [1] 0.006209665
```

In any case, the probability to get the observed value by chance is 0.0062 meaning the null hypothesis can be rejected with $p < 0.01$ and this measurement does not support the hypothesis that $\mu = 5$ in the population.

Paired Z-test can be used to compare two independent samples in a similar way. In this case, we are interested in the difference of the means ($H_0 : \mu_1 = \mu_2$) normalized by the combined SEM i.e.

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

The interpretation of this value is the same as above: a Z-score close to zero supports the null hypothesis and p-values can be directly estimated using the CDF.

In this form, the variances of the populations must be known and this is not always true. However, they can be replaced with sample variances (s_1^2 and s_2^2) in case sample sizes are large enough. Depending on the source, limiting values of 30 or 50 for n_1 and n_2 are used. In practice, the t-test is more useful for small samples.