An (incomplete) summary for Business Statistics 2014/2015 (March 2015)

Notice

This summary is not part of the official course material. Everything written here can also be found in the book, in the slides, or in the tutorial exercises. SPSS is not discussed here and also chapters 1-4 are not (yet) discussed here.
This summary may be helpful to have next to the slides, book, and exercises. It is, however, by no means self-supporting and one should not try to learn any statistics by reading just a summary.
The latest version is maintained at:
http://personal.vu.nl/r.a.sitters/DoaneSummary.pdf

Some notation

The parameters of the population are usually denoted by Greek letters: $\mu, \sigma, ...$ Their estimators are denoted by upper case Roman letters: $\overline{X}, S$, ... and the values of these estimators (the estimates) by lower case Roman letters: $\bar{x}, s$, ... However, for the population parameter for proportion (π), the book uses the lower case p to denote both the estimator X/n as well as its value x/n. Actually, upper and lower case notation is not consistent in the book. It is good to understand the difference but don't worry too much about which to use.

Sometimes μ_x, σ_x is written instead of μ, σ to emphasize that it refers to the variable X. Sometimes π_0, μ_0, etc. is written instead of π, μ to emphasize that it refers to the value of the parameter in H_0. For example, if $H_0 : \mu \leqslant 2$ then $\mu_0 = 2$.

In this summary, 'Ex.' stands for 'Example' and 'e.g.' means 'for example'.

Chapter 5: Probability

Example used below: Throwing a dice once.
Definitions:

- *Sample space* S: The set of all outcomes ($S = \{1, 2, 3, 4, 5, 6\}$)
- *Simple event*: Each of the outcomes ($1, 2, 3, 4, 5$ and 6)
- *Event* A: Subset of S (For example, $A = \{1, 2, 4\}$)
- *Mutually exclusive events*: Having empty intersection. (Ex., $A = \{1, 3\}, B = \{2, 6\}$)
- *Collectively exhaustive events*: Together covering S: (Example, $A = \{1, 2, 3, 4\}, B = \{4, 5, 6\}$. Not mutually exclusive since 4 is in both.)

Notation:

- $P(A)$: *Probability* of event A. Ex. $P(A) = 1/2$ if $A = \{1, 2, 3\}$.
- A': The *complement* of event A. Hence, $P(A) + P(A') = 1$.
- $A \cap B$: *Intersection* of A and B. Ex. If $A = \{1, 2\}$ and $B = \{2, 3\}$ then $A \cap B = \{2\}$.
- $A \cup B$: *Union* of A and B. Ex. If $A = \{1, 2\}$ and $B = \{2, 3\}$ then $A \cup B = \{1, 2, 3\}$.

More definitions:

- *Conditional probability*: $P(A|B) = \frac{P(A \cap B)}{P(B)}$: the probability of A given that B is true.
- Events A and B are called *independent* if

$$P(A \cap B) = P(A) \cdot P(B).$$

(Equivalent: A and B independent if $P(A|B) = P(A)$ or if $P(B|A) = P(B)$.)

More events: Events $A_1, A_2, \ldots, A_n$ are called independent if

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1) \cdot P(A_2) \cdot \cdots \cdot P(A_n).$$

▷ Bayes' theorem (*Not for 2015*)

- Bayes' theorem:
$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}.$$

- Extended form:
$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|B') \cdot P(B')}.$$

- General form:
If $B_1, B_2, \ldots, B_n$ are mutually exclusive and collectively exhaustive then for any i,
$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{P(A|B_1) \cdot P(B_1) + \cdots + P(A|B_n) \cdot P(B_n)}.$$

▷ Counting

- $n! = n(n-1)(n-2)\ldots 1$. (The number of ways to order n items.)

- ${}_nP_r = \frac{n!}{(n-r)!} = n(n-1)\ldots(n-r+1)$.
(The number of ways to choose r ordered items out of n items.)

- ${}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$.
(The number of ways to choose r unordered items out of n items.)

Chapter 6: Discrete Probability Distributions

▶ Some definitions

- A *random variable* is a function or rule that assigns a numerical value to each outcome in the sample space of a random experiment. Random variables are also called stochastic variables.
- A *discrete* random variable has a countable number of distinct values, e.g, $0, 1, 2, 3, \ldots$
- Upper case letters are used to represent random variables (e.g., X, Y).
- Lower case letters are used to represent values of the random variable (e.g., x, y).
- A *discrete probability distribution* assigns a probability to each value of a discrete random variable X.
- Notation: $P(X = x_i)$, or simply $P(x_i)$, is the probability that X takes value x_i.

Expected value (various notation):

$$E(X) = \mu_x = \mu = \sum_{i=1}^{n} x_i P(x_i). \tag{1}$$

Variance:

$$\texttt{var}(X) = V(X) = \sigma_x^2 = \sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 P(x_i). \tag{2}$$

Standard deviation:

$$\sigma = \sqrt{\sigma^2}.$$

▶ Some distributions

▷ Uniform distribution (discrete): $X \sim Uniform(a, b)$.

All simple events have the same probability. Here, we assume that X only takes the values $a, a+1, \ldots, b$.

$$P(X = x) = \frac{1}{b - a + 1}, \quad \mu = \frac{a + b}{2}, \quad \text{and } \sigma = \sqrt{\frac{(b - a + 1)^2 - 1}{12}}.$$

▷ Bernoulli distribution: $X \sim Bernoulli(\pi)$.

The parameter π is the probability of success. If success, then we give variable X the value 1. Otherwise it is zero.

$$P(X = 1) = \pi, \text{ and } P(X = 0) = 1 - \pi.$$

From the definition of μ and σ above it follows directly that

$$\mu = \pi \quad \text{and} \quad \sigma = \sqrt{\pi(1-\pi)}.$$

N.B. For all these common distributions, the formulas for μ ad σ are well-known so we do not have to use the definitions (1) and (2) to compute them. We simply take the precalculated formulas. As an exercise, let us check the formulas for once here:

$$\mu = \sum_{i=1}^{n} x_i P(x_i) = 0(1-\pi)+1\pi = \pi \quad \textit{and} \quad \sigma^2 = \sum_{i=1}^{n} (x_i-\mu)^2 P(x_i) = (0-\pi)^2(1-\pi)+(1-\pi)^2\pi = \pi(1-\pi).$$

▷ Binomial distribution: $X \sim Bin(n, \pi)$.

When we repeat a Bernoulli experiment n times and count the number X of successes we get the Binomial distribution.

$$P(X = x) = \binom{n}{x} \pi^x (1-\pi)^{n-x} = \frac{n!}{x!(n-x)!} \pi^x (1-\pi)^{n-x}.$$

$$\mu = n\pi \quad \text{and} \quad \sigma = \sqrt{n\pi(1-\pi)}.$$

▷ Poisson distribution: $X \sim Poisson(\lambda)$.

Often applied in an arrival process. The parameter λ is the average number of arrivals per time unit. (Other examples: number of misprints per sheet of paper; number of errors per new car.)

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \mu = \lambda, \quad \sigma = \sqrt{\lambda}.$$

Extra (not in book and you may skip this remark): The Poisson distribution can be derived from the Binomial distribution. Consider a time horizon of length T and consider a subinterval I of length 1. Place $n = \lambda T$ points (the arrivals) at random on the large interval. For any of these points, the probability that it will end up in subinterval I is $1/T$. If X is the number of points in I then $X \sim Bin(n, \pi)$ with $n = \lambda T$ and $\pi = 1/T$. Note that $\mu_X = n\pi = \lambda$. If $T \to \infty$ then $X \sim Poisson(\lambda)$. This also shows that the Binomial distribution can be approximated by the Poisson distribution when n is large and π is small.

▷ Hypergeometric distribution: $X \sim Hyper(N, n, S)$.

Similar to the Binomial but now *without* replacement. There are N items and S of these are called a 'success'. If we take 1 item at random then the probability of success is S/N. We repeat this n times and let X be the number of successes. If the item is placed back

(or replaced) each time then $X \sim Bin(n, \pi = \frac{S}{N})$. If we do *not* replace the item then we get the hypergeometric distribution.

$$P(X = x) = \frac{\binom{S}{x}\binom{N-S}{N-x}}{\binom{N}{n}}.$$

Denote $\pi = \frac{S}{N}$. Then,

$$\mu = n\pi, \quad \sigma = \sqrt{n\pi(1-\pi)}\sqrt{\frac{N-n}{N-1}}.$$

Note that μ ad σ are the same as for the Binomial distribution except for the so called *finite population correction factor* $\sqrt{\frac{N-n}{N-1}}$.

▷ **Geometric distribution:** $X \sim Geometric(\pi)$**.**

Repeat the Bernoulli experiment until the first success. Let X be the number of trials needed. Then,

$$P(X = x) = (1-\pi)^{x-1}\pi \quad \text{and} \quad P(X \leqslant x) = 1 - (1-\pi)^x.$$

$$\mu = \frac{1}{\pi}, \quad \sigma = \sqrt{\frac{1-\pi}{\pi^2}}.$$

▶ Addition and multiplication of random variables.

Let X be a random variable, and a, b be arbitrary numbers. Then, $aX + b$ is again a random variable and has mean

$$\mu_{aX+b} = a\mu_X + b.$$

If $a \geqslant 0$ then the standard deviation of $aX + b$ is

$$\sigma_{aX+b} = a\sigma_X.$$

Let X and Y be random variables. Then

$$\mu_{X+Y} = \mu_X + \mu_Y \quad \text{and} \quad \sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2 + 2\sigma_{X,Y}} \quad ,$$

where $\sigma_{X,Y}$ is the *covariance* between X and Y. If X and Y are independent then $\sigma_{X,Y} = 0$. In that case, $\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$.

(NB. The book denotes the covariance by σ_{XY} (i.e., without the comma). But that is a bit confusing since it may be mistaken for the standard deviation of the product XY, which is not the same as the covariance.)

▶ Approximating one distribution by another (discrete → discrete).

We use the following rules of thumb for approximating one distribution by another.

- Hyper → Bin

  If we take n items from a population of N items then it does not really matter if we do this with or without replacement when N is much larger than n:

  $X \sim Hyper(N, n, S) \longrightarrow X \sim Bin(n, \pi), \quad$ with $\pi = \frac{S}{N}$, if $n/N < 0.05$.

- Bin → Poisson

  The binomial distribution approaches the Poisson distribution when n is large and π is small.

  $X \sim Bin(n, \pi) \quad \longrightarrow X \sim Poisson(\lambda)$, with $\lambda = n\pi$, if $n \geqslant 20$ and $\pi \leqslant 0.05$.

Chapter 7: Continuous Probability Distributions

▶ Some definitions

A continues probability distribution is described by its probability density function (PDF), $f(x)$. Properties of $f(x)$:

- $f(x) \geqslant 0$ for all x.
- The total area under the curve function is 1 ($\int_{-\infty}^{\infty} f(x)\ dx = 1$).
- $P(a \leqslant X \leqslant b)$ is equal to the area under the curve between a and b ($= \int_a^b f(x)\ dx$).
- There is no difference between $<$ and $\leqslant$ or between $>$ and $\geqslant$: $P(X \geqslant b) = P(X > b)$.

For any random variable X:

$$\mu_x = \int_{-\infty}^{\infty} x f(x)\ \ dx \text{ and } \ \sigma_x^2 = \int_{-\infty}^{\infty} (x - \mu_x)^2 f(x)\ dx. \tag{3}$$

Note the following remark from the book: '... in this chapter, the means and variances are presented without proof ...' That means, you do not need to worry about the integrals (3) *above. We won't use them since we simply take the precalculated answer as a starting point.*

▶ Some distributions

▷ Uniform distribution (continuous): $X \sim Uniform(a, b)$.

The probability density function is constant between a and b and zero otherwise.

$$f(x) = \frac{1}{b-a} \text{ (for } a \leqslant x \leqslant b), \quad \mu = \frac{a+b}{2}, \quad \sigma = \sqrt{\frac{(b-a)^2}{12}}.$$

▷ Normal distribution: $X \sim N(\mu, \sigma)$.

Also called the Gaussian distribution. The domain is $[\infty, \infty]$. The density function is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}((x-\mu)/\sigma)^2}$. This function will not be used directly here but it follows from this function that, if $X \sim N(\mu, \sigma)$ then

$$\mu_x = \mu \ \text{ and } \ \sigma_x = \sigma.$$

If $\mu = 0$ and $\sigma = 1$ then we call this the *standard* normal distribution: $N(0, 1)$. Any normal distribution can simply be transformed into a standard normal distribution:

$$\text{If } \ X \sim N(\mu, \sigma) \ \text{ then } \ Z = \frac{X - \mu}{\sigma} \sim N(0, 1). \tag{4}$$

For a normal distribution (See also the empirical rule of Chapter 4):

- $P(\mu - \ \sigma \leqslant X \leqslant \mu + \ \sigma) \approx 0.68$ (68% change of being at most σ away form μ.)
- $P(\mu - 2\sigma \leqslant X \leqslant \mu + 2\sigma) \approx 0.95$ (95% change of being at most 2σ away form μ.)
- $P(\mu - 3\sigma \leqslant X \leqslant \mu + 3\sigma) \approx 0.997$ (99.7% change of being at most 3σ away form μ.)

▷ **Exponential distribution:** $X \sim Exp(\lambda)$.

Closely related to the Poisson distribution. If X is the arrival time between two customers and $X \sim Exp(\lambda)$ then the number of customers which arrive in one unit of time has a Poisson distribution with parameter λ. The density function is $f(x) = \lambda e^{-\lambda x}$, for $x \geqslant 0$.

$$P(X \leqslant x) = 1 - e^{-\lambda x}, \quad P(X \geqslant x) = e^{-\lambda x}, \quad \mu = 1/\lambda \quad \text{and } \sigma = 1/\lambda.$$

▷ **More distributions**

Students t, F, $\chi^2, \ldots$

▶ Approximating one distribution by another. (Discrete → Normal)

- Bin → Normal

  The binomial distribution approaches the normal distribution when n increases (and π remains fixed). The rule of thumb is:

  $X \sim Bin(n, \pi) \to X \sim N(\mu, \sigma)$, with $\mu = n\pi$, and $\sigma = \sqrt{n\pi(1-\pi)}$ if $n\pi \geqslant 5$ and $n(1-\pi) \geqslant 5$.

  Note that the **book uses** $\geqslant 10$ here instead of $\geqslant 5$. When we approximate the binomial by a normal then a **continuity correction** should be made.

  Example: $X \sim Bin(\pi = 0.5, n = 20)$. Then, (using the exact binomial distribution) $P(X \leqslant 10) \approx 0.5881$. Using the normal approximation *without* correction gives:

  $$P(X \leqslant 10) \approx P(Z \leqslant (10 - 10)/\sqrt{5}) = 0.5.$$

  Using the normal approximation *with* correction gives a much closer result:

  $$P(X \leqslant 10) \approx P(Z \leqslant (10\frac{1}{2} - 10)/\sqrt{5}) = 0.5885.$$

- Poisson → Normal

  The Poisson distribution approaches the normal distribution when $\lambda \to \infty$. The table goes up to $\lambda = 20$ and we should use the table if possible. The approximation works fine for $\lambda \geqslant 10$ though.

  $X \sim Poisson(\lambda) \to X \sim N(\mu, \sigma)$, with $\mu = \lambda$, and $\sigma = \sqrt{\lambda}$ for $\lambda \geqslant 10$.

Chapter 8: Sampling Distributions and Estimation

- *Sample statistic*: Random variable as function of the sample. *Ex.*: $Y = X_1 + \cdots + X_n$.
- *Estimator*: Sample statistic which estimates the value of a population parameter. *Ex.*: $\overline{X}$.
- *Estimate*: Value of the estimator in a particular sample. *Ex.*: $\bar{x} = 9.3$
- *Sampling error*: Difference between the estimate and the population parameter. *Ex.*: $\bar{x} - \mu$.

Some important estimators:

- *Sample mean*: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (the random variable), or $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the estimate).
- *Sample variance*: $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ (the random variable) or $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ (estimate).
- *Sample standard deviation* $S = \sqrt{S^2}$ and $s = \sqrt{s^2}$.

One might wonder why we divide by $n-1$ in the formula for S^2 and not by n as is done in the formula for σ in Chapter 4. The formula above is an unbiased estimator for σ^2, that means, $E[S^2] = \sigma^2$.

▶ Distribution of the mean

One of the most important properties to remember when taking samples is that the expected value of the sample mean is the same as the mean of the population distribution but the variance reduces by a factor $\sqrt{n}$.

For *any* distribution X:

$$\mu_{\bar{x}} = \mu_x \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}.$$

- The value $\sigma_x/\sqrt{n}$ is called the *standard error of the mean* (SE)

In particular, if $X \sim N(\mu, \sigma)$ then $\overline{X} \sim N(\mu, \sigma/\sqrt{n})$. Further, for any arbitrary distribution X, the distribution of $\overline{X}$ approaches a normal distribution for $n \to \infty$. This gives the Central Limit Theorem (CLT) for the mean:

Central Limit Theorem (CLT)

For large enough n (see (@) below):

$$\overline{X} \sim N(\mu_x, \frac{\sigma_x}{\sqrt{n}})$$

and therefore (see (4))

$$\frac{\overline{X} - \mu_x}{\sigma_x/\sqrt{n}} \sim N(0,1).$$

What is 'large enough'? We use the following rule of thumb :

(@)
- If X is normally distributed then any n is fine.
- If X is symmetrically distributed (without extreme outliers) then $n \geqslant 15$ is fine.
- If X is not symmetric then we need $n \geqslant 30$.

▷ Students t-distribution

Often, σ is not known and we use the sample esitimate S instead. If X has a normal distribution then the distribution of $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ does not depend on μ and σ. It is called the *students t-distribution* with $n-1$ degrees of freedom (d.f.) t_{n-1}.
For arbitrary distribuation X we can use the rule of thumb above and reach the same conclusion:

For large enough n (see (@))

$$\frac{\overline{X} - \mu_x}{S/\sqrt{n}} \sim t_{n-1}.$$

We say that the *sample statistic* $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ has the *sampling distribution* t_{n-1}.

Properties of the t-distribution:

- The tails are 'fatter' than that of the normal distribution.
- The t-distribution approaches the $N(0,1)$ distribution for $d.f. \to \infty$.

▷ Confidence interval for μ

We can only make a confidence interval for μ if we can safely assume that $\overline{X}$ is approximately normally distributed and that holds if condition (@) is satisfied.

- Confidence interval for μ (σ known):

$$\bar{x} - z\frac{\sigma}{\sqrt{n}} \;<\; \mu \;<\; \bar{x} + z\frac{\sigma}{\sqrt{n}}, \text{ where } z = z_{\alpha/2}.$$

- Confidence interval for μ (σ not known):

$$\bar{x} - t\frac{s}{\sqrt{n}} \;<\; \mu \;<\; \bar{x} + t\frac{s}{\sqrt{n}}, \text{ where } t = t_{n-1;\alpha/2}$$

- For finite populations (size N) there is an additional *finite population factor*, $\sqrt{\frac{N-n}{N-1}}$:

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \text{ (known } \sigma) \qquad \text{and} \qquad \bar{x} \pm t\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \text{ (unknown } \sigma).$$

▶ Distribution of the proportion

Assume that a fraction π of the population has a certain property. For example, if 40% of the people travel by train then the *population proportion* is $\pi = 0.4$. If we take a sample of size n and let X be the number of successes (people traveling by train) in the sample then the *sample proportion* is:

$$p = \frac{X}{n}.$$

What is the distribution of p? Remember what we did for the distribution of $\bar{X}$. There we concluded that for large enough n (see (@)) the distribuation is approximately normal. One problem was that the distribution depends on σ and that was solved by using the sample standard deviation s instead of σ. Here, the conclusion is similar: For large enough n, the distribution of p is approximately normal. (See (7) below). Note that the distribution depends on π which may be unknown and this is problemetic if, for example, we want to make a confidence interval for π. We will solve this by simply using p instead of π. In short, that is how it is explaind in the book. However, a lot more can said about the distribution of p since computations on p can be done by using $p = X/n$ and then doing the computation on X instead. Although this is actually not new and follows from what we learnt about the Binomial distribuation, it may be good to list it here. (See also the exercises of week 2) The distribution to use depends on the (relative) size of poulation and sample.

- **Infinite population** (See for an example Q14, tutorial 2)
  More precisely, the population size is large compared to the size of the sample. In that case, we can treat it as taking samples with replacement. Then,

$$X \sim Bin(n, \pi).$$

  We distinguish between small and large samples.

  - For a small sample we can use this exact binomial distribution of X.
  - If the sample is relatively large (to be precise, if $n\pi \geqslant 5$ and $n(1-\pi) \geqslant 5$) then the binomial can be approximated by a normal distribution: $X \sim N(\mu_x, \sigma_x)$, where

$$\mu_x = n\pi \quad \text{and} \quad \sigma_x = \sqrt{n\pi(1-\pi)}.$$

Thus,

$$Z = \frac{X - \mu_x}{\sigma_x} = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \sim N(0,1). \tag{5}$$

Equivalently, we may work with p instead of X (as is done in the book). Substituting $X = np$ implies

$$\mu_p = \pi \quad \text{and} \quad \sigma_p = \sqrt{\pi(1-\pi)/n}, \tag{6}$$

and

$$Z = \frac{p - \mu_p}{\sigma_p} = \frac{p - \pi}{\sqrt{\pi(1-\pi)/n}} \sim N(0,1). \tag{7}$$

NB. Although the statistics (5) and (7) are equivalent, using (5) has the advantage that a continuity correction can be made.

Example:
Let $\pi = 0.5$ and $n = 20$ and assume our sample has $p = 0.4$, i.e. $X = 8$. Say we want to compute $P(p \leqslant 0.4)$. Then,

* No continuity correction:
  $P(p \leqslant 0.4) \approx P(Z \leqslant \frac{0.4-0.5}{\sqrt{0.5(1-0.5)/20}}) = P(Z \leqslant -0.89) = 0.19.$
* Using X and the continuity correction:
  $P(p \leqslant 0.4) = P(X \leqslant 8) \approx P(Z \leqslant \frac{8\frac{1}{2}-10}{\sqrt{20 \cdot 0.5(1-0.5)}}) = P(Z \leqslant -0.67) = 0.2511.$
* Compare this with the true binomial probability: $P(X \leqslant 8) = 0.2517.$

- **Finite population size** N (See for an example Q15, tutorial 2).
  In this case,

  $$X \sim Hyper(N, n, S), \text{ with } S = N\pi.$$

  Again, we distinguish between small and large samples.

  - For a small sample we can use this exact hypergeometric distribution of X.
  - If the sample is relatively large then we can use a normal approximation to the hypergeometric distribution. 'Relatively large' in this case means $n\pi \geqslant 5, n(1-\pi) \geqslant 5$ and, additionally, $(N-n)\pi \geqslant 5$ and $(N-n)(1-\pi) \geqslant 5$. In that case, $X \sim N(\mu_x, \sigma_x)$, where

    $$\mu_x = n\pi \quad \text{and} \quad \sigma_x = \sqrt{n\pi(1-\pi)}\sqrt{\frac{N-n}{N-1}}.$$

    Thus,

    $$Z = \frac{X - \mu_x}{\sigma_x} = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}\sqrt{\frac{N-n}{N-1}}} \sim N(0,1). \tag{8}$$

Equivalently, we may work with p instead of X by substituting $X = np$. (But that makes a continuity correction impossible.) In that case,

$$\mu_p = \pi \quad \text{and} \quad \sigma_p = \sqrt{\pi(1-\pi)/n}\sqrt{\frac{N-n}{N-1}}, \tag{9}$$

and

$$Z = \frac{p-\mu_p}{\sigma_p} = \frac{p-\pi}{\sqrt{\pi(1-\pi)/n}\sqrt{\frac{N-n}{N-1}}} \sim N(0,1). \tag{10}$$

▷ Confidence interval for proportion π

We only consider the case of large samples. In that case,

$$Z = \frac{p-\pi}{\sigma_p} \sim N(0,1).$$

Then, the confidence interval for π becomes: $p \pm z_{\alpha/2}\sigma_p$. Again, we distinguish between large (say infinite) populations, and finite populations (of size N). In both cases σ_p is a fuction of π (See (6) and (9)). However, since we are making a confidence interval for π, it is fair to assume that π is not already known. In that case we simply replace π by p. For infinite populations the condition becomes $np \geqslant 5$ and $n(1-p) \geqslant 5$ and the *confidence interval* for π becomes

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

For finite populations (size N) the condition is $np \geqslant 5, n(1-p) \geqslant 5, (N-n)p \geqslant 5$ and $(N-n)(1-p) \geqslant 5$ and the *confidence interval* for π becomes

$$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}.$$

► Distribution of the variance.

For normal distributions, the distribution of $(n-1)s^2/\sigma^2$ depends on n only and is called the Chi-Square distribution:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

▷ Confidence interval for variance σ^2.

$$\frac{(n-1)s^2}{\chi^2_U} \;<\; \sigma^2 \;<\; \frac{(n-1)s^2}{\chi^2_L}.$$

Here, χ^2_U and χ^2_L are the *Upper* and *Lower* critical values for the corresponding value of α, which can be derived from the table in the book. (Note the switch: The upper value is left and the lower value is on the right).

▶ Sample size estimation.

The *margin of error* (E) is half the size of the confidence interval. Given E, what should be the sample size n to obtain a confidence interval with margin of error at most E? If the computed n is not a whole number then round up.

▷ Sample size estimation for μ

- Size for known (or estimated) σ:

$$E = z\frac{\sigma}{\sqrt{n}} \quad \rightarrow \quad n = \left(\frac{z\sigma}{E}\right)^2. \tag{11}$$

- Size for unknown σ:

$$E = z\frac{s}{\sqrt{n}} \quad \rightarrow \quad n = \left(\frac{zs}{E}\right)^2.$$

  There is something strange with the formula above. Looking at the formula for the confidence interval it should actually be $E = t\frac{s}{\sqrt{n}}$, where $t = t_{n-1;\alpha/2}$. However, it is impossible to compute n from this since n is a parameter of t. Therefore, we simply use $z = z_{\alpha/2}$. The other thing to notice is that the sample variation s is probably not known since we are computing the size n of the sample which we still need to take. In that case, work as follows:

- Size for both σ and s unknown. Use (11) and do one of the following:
  - Take a preliminary (small) sample and use its sample standard deviation for σ.
  - If distribution is uniform then use $\sigma = \sqrt{(b-a)/12}$, where all values are assumed to lie between a and b.
  - If distribution is (approximately) normal use $\sigma = (b-a)/6$ (or more conservative $\sigma = (b-a)/4$).
  - For Poisson arrivals we have $\sigma = \sqrt{\lambda}$. Make a guess for the arrival rate λ.

▷ Sample size estimation for proportion π

- Size for known (or estimated) value π:

$$E = z\sqrt{\frac{\pi(1-\pi)}{n}} \quad \rightarrow \quad n = z^2\frac{\pi(1-\pi)}{E^2}. \tag{12}$$

  However, π is probably not known since the goal is to construct an interval for π. Use p instead:

- Size for known value p:

$$E = z\sqrt{\frac{p(1-p)}{n}} \quad \rightarrow \quad n = z^2\frac{p(1-p)}{E^2}.$$

  However, the sample proportion p is probably not known/given either since we are computing the size n of the sample which we still need to take. In that case, work as follows:

- Size for π and p unknown. Use (12) and do one of the following: In that case we could do one of the following:
  - Take a preliminary (small) sample and use its proportion for π.
  - Take a preliminary (small) sample but instead of using p directly, build a confidence interval for π and use the border which is nearest to 0.5
  - Use historical data to estimate π.
  - Take $\pi = 0.5$. This is the conservative approach and gives the largest value n.

Chapter 9: One-Sample Hypothesis

▶ Testing

H_0: Null hypothesis.
H_1: Alternative hypothesis. (= what we want to show to be probably true.)

▷ Errors.

- Type I error: Rejecting a true H_0. Probability(Type I error)= α. Value α is chosen.
- Type II error: Not rejecting a false H_0. Probability(Type II error)= β.
  Value β is *not* chosen but follows from α, n, H_0, and the (unknown) real distribution.

The *power of the test* is the probability of rejecting a false H_0 (which is not an error but is what we want). Power $= 1 - \beta$.
We want α and β both to be small but reducing α will increase β if we make no other changes. Increasing the sample size is a way to decrease both α and β.

▷ 5 steps for testing:

1. Hypothesis + α
2. Sample statistic and indicate rejection region.
3. Test statistic + its (approximate) distribution. Assumptions and/or conditions.
4. Value of test statistic + critical value(s) or give p-value.
5. Make a decision (Reject/Do not reject Ho) + conclusion in words.

In some tests, like ANOVA and regression one should state the model as a step 0.

▶ Test for the population mean μ

$$\begin{array}{lll} H_0 : \mu \geqslant \mu_0 & & H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 & \text{or} & H_1 : \mu \neq \mu_0 \end{array} \quad \text{or} \quad H_0 : \mu \leqslant \mu_0 \quad H_1 : \mu > \mu_0$$

Test statistic (σ known):

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Test statistic (σ unknown):

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}.$$

Confidence interval for μ (See Chapter 8):

$$\bar{x} \pm t \cdot s/\sqrt{n}, \quad \text{with} \quad t = t_{n-1;\alpha/2}.$$

A two-sided test for μ is similar to a confidence interval:

$$H_0 : \mu = \mu_0 \text{ rejected} \quad \Longleftrightarrow \quad \mu_0 \text{ outside confidence interval.}$$

In all cases above we assume that the condition (@) (see Chapter 8) holds.

▶ Test for the population proportion π

$H_0 : \pi \geqslant \pi_0$		$H_0 : \pi = \pi_0$		$H_0 : \pi \leqslant \pi_0$
$H_1 : \pi < \pi_0$	or	$H_1 : \pi \neq \pi_0$	or	$H_1 : \pi > \pi_0$

Instead of (@) we use the condition: $n\pi_0 \geqslant 5$ and $n(1-\pi_0) \geqslant 5$.
Test statistic (book version):

$$Z = \frac{p - \pi_0}{\sigma_p} = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}.$$

Let X be the number of successes: $X = pn$. Replacing p by X/n gives an equivalent statistic:

$$Z = \frac{X - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}.$$

The two statistics above are identical. However, the latter has the advantage that an additional continuity correction is possible. (See the section on proportion.) Continuity correction is not really needed if you notice that the p-value is much smaller or much larger than α. (The continuity correction will not change the p-value a lot.) However, if the p-value is close to α then continuity correction might make a difference and should therefore be made in that case.

If condition $n\pi_0 \geqslant 5, n(1-\pi_0) \geqslant 5$ fails then use the exact distribution: $X \sim Bin(n, \pi_0)$.

Example: Let $n = 4, X = 3$. $H_0 : \pi \leqslant 0.2$. Then, p-value=$P(X \geqslant 3) = P(X = 3) + P(X = 4) = 0.026 + 0.002 = 0.028$.

▶ Test for the population variance.

$H_0 : \sigma^2 \geqslant \sigma_0^2$		$H_0 : \sigma^2 = \sigma_0^2$		$H_0 : \sigma^2 \leqslant \sigma_0^2$
$H_1 : \sigma^2 < \sigma_0^2$	or	$H_1 : \sigma^2 \neq \sigma_0^2$	or	$H_1 : \sigma^2 > \sigma_0^2$

Sample statistic: S^2. Test statistic:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}.$$

Assumption: Population is normally distributed.

Look up critical values, χ^2_L, χ^2_U, from table. Reject one or two sides, depending on H_0.

Chapter 10: Two-Sample Hypothesis

For any two stochastic variables X_1, X_2:

$$\mu_{\overline{X}_1-\overline{X}_2} = \mu_1 - \mu_2 \tag{13}$$

If X_1 and X_2 are independent (as we assume here) then $\texttt{cov}(\overline{X}_1, \overline{X}_2) = 0$. In that case

$$\begin{aligned} \texttt{var}(\overline{X}_1 - \overline{X}_2) &= \texttt{var}(\overline{X}_1) + \texttt{var}(\overline{X}_2) - 2\texttt{cov}(\overline{X}_1, \overline{X}_2) \\ &= \texttt{var}(\overline{X}_1) + \texttt{var}(\overline{X}_2). \end{aligned} \tag{14}$$

▶ Difference in means: two independent samples.

$H_0 : \mu_1 - \mu_2 \geqslant 0$ $\quad$ $H_0 : \mu_1 - \mu_2 = 0$ $\quad$ $H_0 : \mu_1 - \mu_2 \leqslant 0$
$H_1 : \mu_1 - \mu_2 < 0$ $\quad$ or $\quad$ $H_1 : \mu_1 - \mu_2 \neq 0$ $\quad$ or $\quad$ $H_1 : \mu_1 - \mu_2 > 0$

Instead of testing against 0, we may take any value. For example, we could test $H_0 : \mu_1 - \mu_2 = 10$. In that case $\mu_1 - \mu_2 = 10$ in the formulas below. Usually, the benchmark is 0 and in that case $\mu_1 - \mu_2 = 0$ in the formalas below.

The choice of the test statistic follows from the next three possible cases. The CLT conditions (@) (See Chapter 8) must hold for both samples.

1. σ_1, σ_2 known.
2. σ_1, σ_2 unknown but assumed equal.
3. σ_1, σ_2 unknown and not assumed equal.

1. Use test statistic
$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sigma_{\overline{X}_1-\overline{X}_2}}, \quad \text{where } \sigma_{\overline{X}_1-\overline{X}_2} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}.$$
2. Use pooled variance t-test.
$$t = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{s_p^2/n_1 + s_p^2/n_2}} \sim t_{n_1+n_2-2}, \text{ where } s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}.$$
3. Use separate variance t-test.
$$t = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \sim t_{df},$$
where df = some large formula. (See book or formula sheet). Round down.
A quick (but not very accurate!) rule is to use $df \approx \min\{n_1 - 1, n_2 - 1\}$.

If $n_1 = n_2$ then the t-values for case 2 and case 3 are equal but df may differ.

Confidence intervals:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2,\alpha/2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \quad \text{(pooled)}, \qquad (\bar{x}_1 - \bar{x}_2) \pm t_{df,\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}} \quad \text{(separate)}.$$

▶ Difference in means: paired samples.

Let

$$d = X_1 - X_2.$$

Then, d is a stochastic variable with $\mu_d = \mu_1 - \mu_2$. (You may also see D instead of d but the book uses the lower case.) A test for $\mu_1 = \mu_2$ now simply becomes a one-sample t-test for $\mu_d = 0$.

$$\begin{array}{lll} H_0 : \mu_d \geqslant 0 & & \\ H_1 : \mu_d < 0 & \text{or} & \end{array} \qquad \begin{array}{lll} H_0 : \mu_d = 0 & & \\ H_1 : \mu_d \neq 0 & \text{or} & \end{array} \qquad \begin{array}{lll} H_0 : \mu_d \leqslant 0 & & \\ H_1 : \mu_d > 0 & & \end{array}$$

Let $\bar{d}$ be the sample mean of d and s_d be the sample standard deviation of d. Then, the test statistic for this one sample t-test is:

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \sim t_{n-1}.$$

Assumption: The condition (@) should hold for d.

NB. It is also possible to test for another value than 0. For example: $H_0 : \mu_d \geqslant -8$ (See Q2 tutorial 4). In that case just take $\mu_d = -8$ in the formula for t above.

Confidence interval for μ_d:

$$\bar{d} \pm t_{n-1;\alpha/2}\frac{s_d}{\sqrt{n}}.$$

▶ Difference in proportions (independent samples).

First, let us compute the mean and varaiance of the difference. From (6),(13),(14) we see that:

$$\begin{aligned} \mu_{p_1-p_2} &= \pi_1 - \pi_2 \\ \texttt{var}(p_1 - p_2) &= \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}. \end{aligned}$$

The possible hypotheses are:

$$\begin{array}{lll} H_0 : \pi_1 - \pi_2 \geqslant D_0 & & H_0 : \pi_1 - \pi_2 = D_0 & & H_0 : \pi_1 - \pi_2 \leqslant D_0 \\ H_1 : \pi_1 - \pi_2 < D_0 & \text{or} & H_1 : \pi_1 - \pi_2 \neq D_0 & \text{or} & H_1 : \pi_1 - \pi_2 > D_0 \end{array}$$

Two test statistics are distinguished depending on D_0 being zero or not.

- $D_0 = 0$. In other words: we assume the proportions are the same. In that case we use a pooled proportion for the estimate: $\bar{p} = \frac{x_1+x_2}{n_1+n_2}$.
  Test statistic:
  $$z = \frac{p_1 - p_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} \sim N(0,1). \tag{15}$$
  Conditions: $n_1\bar{p} \geqslant 5, n_1(1-\bar{p}) \geqslant 5$ and $n_2\bar{p} \geqslant 5, n_2(1-\bar{p}) \geqslant 5$.

- $D_0 \neq 0$. Test statistic:
  $$z = \frac{p_1 - p_2 - D_0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1).$$
  Conditions: $n_1 p_1 \geqslant 5, n_1(1-p_1) \geqslant 5$ and $n_2 p_2 \geqslant 5, n_2(1-p_2) \geqslant 5$.

▶ Difference in variances (independent samples)

$$\begin{array}{lll} H_0 : \sigma_1^2/\sigma_2^2 \geqslant 1 & & H_0 : \sigma_1^2/\sigma_2^2 = 1 & & H_0 : \sigma_1^2/\sigma_2^2 \leqslant 1 \\ H_1 : \sigma_1^2/\sigma_2^2 < 1 & \text{or} & H_1 : \sigma_1^2/\sigma_2^2 \neq 1 & \text{or} & H_1 : \sigma_1^2/\sigma_2^2 > 1 \end{array}$$

Test statistic:

$$F = \frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1}. \quad \text{(Reject one or both sides, depending on } H_0\text{.)}$$

Assumption: Both populations are normal.

Critical values F_L, F_R. Only F_R in table. The left critical value can be computed from $F_L = 1/F_R'$, where F_R' is the right critical for labels 1 and 2 reversed. *Tip: Label the samples such that $s_1^2 \geqslant s_2^2$. Then, only F_R is needed.*

The **Levene test** is another test for comparing two variances. The exact test and distribution are not explained in the book. We shall only use its p-value from a given output. Some properties of Levene:
- Normal assumption not needed.
- Always 2-sided ($H_0 : \frac{\sigma_1}{\sigma_2} = 1$) and 1-tailed. Reject high values.
- Also works for comparing more than two variances.

Chapter 11: ANOVA (Comparison of two or more means)

▶ One factor ANOVA

ANOVA is used to test equality of means of more than two samples. Given c samples, we want to test wether $\mu_1 = \mu_2 = \cdots = \mu_c$. When we write $\mu_j = \mu_0 + \alpha_j$ then this is equivalent with testing wether there is a value μ_0 such that $\alpha_1 = \alpha_2 = \cdots = \alpha_c = 0$. In the model, we assume that all populations are normally distributed with the same variance. Let y_{ij} be the i-th observation in the j-th sample (column).

Model:

$$y_{ij} = \mu_0 + \alpha_j + \epsilon_{ij}, \quad \text{with } \epsilon_{ij} \sim N(0, \sigma^2).$$

NB. Doane uses the notation A_j instead of α_j. Further, the book states all formulas in terms of observed values instead of random variables, that means, it uses the lowercase y_{ij} and occasionally uses Y to denote the random variable. On the slides the model is stated in terms of the random variables and use the uppercase notation Y_{ij}. Both are fine. It is good to notice the difference but don't worry about these details.

▷ The assumptions.

As usual, it is assumed that all observations are independent. Other assumptions (which can be seen from the model):

- the populations are normally distributed.
- the populations have equal variances.

ANOVA is quite robust to violation of normality and equal variances. If all groups have equal size then robust against inequality of variances. To test equality of variances, use the Levene from output.

▷ The computation (the ANOVA table).

There are n observations divided over c groups and observation i in group (column) j is denoted by y_{ij}. Further notation (book version):

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij} \text{ (group mean)} \quad \text{and} \quad \bar{y} = \frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n_j} y_{ij} \text{ (overall mean).}$$

- $SSA = \sum_{j=1}^{c} n_j(\bar{y}_j - \bar{y})^2$: Sum of squares among or between groups. Also known as: variation due to factor and explained variation, $d.f. = c - 1$

- $SSE = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$: Error sum of squares. Also known as: Sum of squares within the groups (SSW), variation due to randomness, and unexplained variation, $d.f. = n - c$.
- $SST = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2$: Sum of squares in total, $d.f. = n - 1$

$$SST = SSA + SSE \quad \text{(Total = Among + Error)}$$

The *mean* sum of squares are derived by dividing the totals by their degrees of freedom:

$$MST = \frac{SST}{n-1}, \qquad MSA = \frac{SSA}{c-1}, \qquad MSE = \frac{SSE}{n-c}.$$

▷ The ANOVA test.

In previous sections we have seen three hypothesis ($\geqslant, =, \leqslant$). Here, there is only one hypothesis:

$$\begin{aligned} H_0: &\quad \alpha_1 = \alpha_2 = \cdots = \alpha_c = 0 \\ H_1: &\quad \text{Not all are zero.} \end{aligned}$$

If H_0 is rejected then we speak of a *factor effect* or *treatment effect*. The naive approach would be to do a t-test for each pair. This is not working since it is likely to fail for at least one pair if the number of groups is large (even if all means are equal). The ANOVA test compares all means *simultaneously*.

Test statistic:

$$F = \frac{MSA}{MSE} \sim F_{c-1,n-c}. \quad \text{(reject for large values).}$$

(Under H_0, the expected value of F is 1 and F-values very close to zero are unlikely. However, such a small value is no indication that means are different. So only reject the large values.)

▷ Post-hoc analysis: The Tukey test.

If H_0 is rejected then an additional test is needed to see for which pairs the means differ significantly. This can be done by Tukey's test. (In SPSS, use Tukey's HSD). In this Tukey test, the means of all $c(c-1)/2$ pairs are tested *simultaneously*. For groups j and k this yields:

$$\begin{aligned} H_0: &\quad \mu_j = \mu_k, \\ H_1: &\quad \mu_j \neq \mu_k. \end{aligned}$$

Reject H_0 for pair j, k if

$$|\bar{y}_j - \bar{y}_k| \geqslant T_{c,n-c} \sqrt{MSE \left(\frac{1}{n_j} + \frac{1}{n_k} \right)}, \quad \text{(critical range)}$$

where $T_{c,n-c}$ is a critical value (depending on α) and is derived from a table for Tukey. Note that the righthand side of the inequality is the same for all pairs if all groups have the same size.

▶ Two factor ANOVA without replication (Course: Statistics II)

The levels of the second factor (the rows) are sometimes called the *blocks*. Often, the main factor of interest are the columns and the blockdesign is only added to strengthen the model. Sometimes both factors are of interest. We assume here that there is exactly one observation for each row/column pair. This is the *two factor ANOVA without replication*. With only one observation for each combination, interaction effects cannot be estimated. If there are at least two observations for each row/column pair then we speak of the *two factor ANOVA with replication*. In that case it is useful to look at interactions effects between the two factors. (See next section.)

Let y_{ij} be the observation in row i and column j. Model:

$$y_{ij} = \mu_0 + \alpha_i + \beta_j + \epsilon_{ij}, \quad \text{with } \epsilon_{ij} \sim N(0, \sigma^2).$$

The book uses A and B instead of α and β. Further, note that in one-factor ANOVA, the α_j refers to column j while here we use α_i for row i. That notation is not perfect but it is common so we use it like that here.

As usual, it is assumed that all observations are independent. Other assumptions (which can be seen from the model) are • Normally distributed populations, • Variances are equal.

Ler r, c be the number of rows and columns. We can either test for a row effect or a column effect.

$$\begin{array}{ll} H_0: & \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0 \qquad \text{(Row means are the same.)} \\ H_1: & \text{Not all are zero.} \end{array}$$

$$\begin{array}{ll} H_0: & \beta_1 = \beta_2 = \cdots = \beta_c = 0 \quad \text{(Column means are the same.)} \\ H_1: & \text{Not all are zero.} \end{array}$$

SST	=	total sum of squares
SSA	=	between rows sum of squares (effect of factor A)
SSB	=	between columns sum of squares (effect of factor B)
SSE	=	error sum of squares (residual variation)

$$\text{SST=SSA+SSB+SSE}$$

The *mean* sum of squares, MST,MSA,MSB, and MSE, are derived by dividing by the degrees of freedom. The exact formulas and degrees of freedom are omitted here. They can be found on the formula sheet if needed but often the values can simply be read form

given output. The F-values for the corresponding hypotheses are:

$$F = \frac{MSA}{MSE} \text{ (for rows)}, \qquad F = \frac{MSB}{MSE} \text{ (for columns)}.$$

Tukey test. The Tukey test for the post hoc analysis works the same as for the one factor ANOVA. The test is done for each factor separately but often you will be interested in only one of the two.

▶ Two factor ANOVA with replication (Course: Statistics II)

If there are at least two observations for each row/column pair then we speak of the *two factor ANOVA with replication.* In that case it is useful to look at interaction effects between the two factors. Let y_{ijk} be observation i for row j and column k.

Model:

$$y_{ijk} = \mu_0 + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}, \quad \text{with } \epsilon_{ijk} \sim N(0, \sigma^2).$$

The main effects (row and column) can be tested as was done in the two factor ANOVA without replication. Additionally, we can test for an interaction effect:

$$\begin{aligned} H_0: &\quad \text{All the } (\alpha\beta)_{jk} \text{ are zero.} \\ H_1: &\quad \text{Not all } (\alpha\beta)_{jk} \text{ are zero.} \end{aligned}$$

The interaction sum of squares is denoted by SSI. The equation becomes

$$\text{SST=SSA+SSB+SSI+SSE}$$

The *mean* sum of squares are derived by dividing by the degrees of freedom. The exact formulas and degrees of freedom are omitted here. The F-value for the test on an interaction effect is:

$$F = \frac{MSI}{MSE}.$$

Example

	β_1	β_2	β_3
α_1	10.1	20.2	30.2
	10.2	20.4	30.1
α_2	10.0	20.1	30.5
	10.1	20.4	30.1

	β_1	β_2	β_3
α_1	10.1	20.2	30.2
	10.2	20.4	30.1
α_2	20.0	30.1	40.5
	20.1	30.4	40.1

	β_1	β_2	β_3
α_1	10.1	20.2	30.2
	10.2	20.4	30.1
α_2	30.0	20.1	10.5
	30.1	20.4	10.1

Left: Strong column effect. No row or interaction effect.
Middle: Strong row and column effect. No interaction effect.
Right: No row or column effect but a strong interaction effect.

Chapter 12: Simple linear regression

In *simple linear regression* there is only one x-variable. The next chapter deals with more than one x-variable. We have n pairs of observations (x_i, y_i) coming from populations X and Y. The X is called the *independent* variable and Y is the *dependent* variable. Both are numeric variables and we want to test how Y depends on X. In the regression model, only the Y is considered a *random* variable.
Note the differences with the models from chapter 11 and 15:

ANOVA:	X categorical,	Y numeric,
Regression:	X numeric,	Y numeric.
Chi-square test:	X categorical,	Y categorical,

Some notation: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (average x-value), $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ (average y-value).

▶ Correlation

Correlation measures the degree of linearity between two (random) variables X and Y.

Population correlation coefficient: ρ.

Sample correlation coefficient: $r = \dfrac{SS_{xy}}{\sqrt{SS_{xx}}\sqrt{SS_{yy}}}$, $(-1 \leqslant r \leqslant 1)$, where

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

We can test for ρ being unequal to zero:

$$\begin{aligned} H_0 &: \quad \rho = 0 \quad \text{(No correlation)} \\ H_1 &: \quad \rho \neq 0 \quad \text{(There is correlation).} \end{aligned}$$

Sample statistic: r

Test statistic:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}, \quad \text{(reject both sides)}.$$

A one sided test ($\rho \leqslant 0$, or $\rho \geqslant 0$) is also possible. In that case, reject one side.
A quick (less accurate) test is to reject H_0 if $|r| > 2/\sqrt{n}$.

Note that no test is given in the book for $H_0 : \rho = \rho_0$, with $\rho_0 \neq 0$.

▶ Regression

▷ Regression model

- In the regression model, we assume that (within a certain domain) the data satisfies the following *regression equation*:

$$y = \beta_0 + \beta_1 x + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2) \qquad \text{(assumed model)}.$$

That means, for any observation (x_i, y_i) we have

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where ϵ_i is the *error* for observation i and is assumed to come from a normal distribution $N(0, \sigma^2)$. The parameter β_0 is the *intercept* and β_1 is the *slope*.

- From the sample we try to find estimates b_0, b_1 for the parameters β_0, β_1. The *estimated regression equation* (or *fitted* equation) becomes

$$\hat{y} = b_0 + b_1 x. \qquad \text{(estimated equation)}. \tag{16}$$

In particular, for any x_i we have

$$\hat{y}_i = b_0 + b_1 x_i.$$

The *residual* (or error) for point i is the difference between y_i and the estimate $\hat{y}_i$.

$$e_i = y_i - \hat{y}_i.$$

The estimates b_0, b_1 are found by minimizing $\sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2$. This is called the *ordinary least squares method* (OLS).

▷ Checking the assumptions (LINE)

As can be seen from the regression model, we assume that:

- there is a **L**inear relation between X and Y.

Further, we assume that the error terms ϵ_i

- are **I**ndependent (no autocorrelation),
- are **N**ormally distributed,
- have **E**qual variance (homoscedastic).

The real error terms ϵ_i are not known since we do not know β_0 and β_1. Instead, we use the residuals e_i to check these assumptions.

Independence: Plot the residuals against the X variable and look how the sign (positive or negative) changes. If there is no autocorrelation then a residual e_i will be followed by a residual e_{i+1} of the same sign in roughly half the cases. There is *positive autocorrelation* if positive residuals are mainly followed by positive residuals and negative residuals are mainly followed by negative ones. There is *negative autocorrelation* if positive residuals are mainly followed by negative residuals and negative residuals are mainly followed by positive ones.

Normality: Check skewness and Kurtosis of the residuals. If both between -1 and 1 then OK. Should be OK if $n \geqslant 30$. (Normality violation not a big problem unless there are major outliers.)

Equal variance: Plot the residuals against the X variable. There should be no pattern (such as increasing or decreasing residuals).

▷ The computation (the ANOVA table).

As mentioned, the values b_0 and b_1 are chosen such that the sum of squared residuals SSE is minimized. The following sums of squares are of interest for the analysis:

$SST = \sum_i (y_i - \bar{y})^2$: Total sum of squares, $d.f. = n - 1$.
$SSR = \sum_i (\hat{y}_i - \bar{y})^2$: Regression sum of squares (explained), $d.f. = 1$
$SSE = \sum_i (y_i - \hat{y}_i)^2$: Error/residual sum of squares (unexplained), $d.f. = n - 2$

Property:

$$SST = SSR + SSE. \tag{17}$$

The *mean* sum of squares are derived by dividing the totals by their degrees of freedom:

$$MST = \frac{SST}{n-1}, \qquad MSR = \frac{SSR}{1}, \qquad MSE = \frac{SSE}{n-2}.$$

▷ Testing significance.

The sample statistic b_1 is an estimate for the real slope β_1. If $b_1 \neq 0$ (and it will always be in a random sample) then this does not imply immediately that the real slope β_1 is different from zero. Wether the value is significant depends on the size n of the sample and the variation in Y.

$$\begin{aligned} H_0: &\quad \beta_1 = 0. \\ H_1: &\quad \beta_1 \neq 0. \end{aligned}$$

The test statistic is

$$F = \frac{MSR}{MSE} \sim F_{1,n-2} \qquad \text{(Reject for large values.)}$$

- A large F-value tells you that the best fitting straight line has a slope which is significantly different from zero. Be aware that this is not the same as saying the a straight line fits the data well. Curved data may give high F-values.

▷ Usefulness of the model.

If the slope is significant then the next thing to check is wether the model is useful. This is measured by the R^2 statistic.

The *coefficient of determination* is

$$R^2 = \frac{SSR}{SST}, \quad (0 \leqslant R^2 \leqslant 1).$$

For simple regression (only one X-variable), $R^2 = r^2$. It is the percentage of variation in Y that is explained by the variation in X. If the slope is significant and R^2 is close to 1 then the model is useful. (Figure 1-A). Be aware that a significant slope together with a small value of R^2 (Figure 1-B) may still be useful for some applications. Also, insignificant slope may be due to a too small sample (Figure 1-C) and the model could be useful after all but we just can't tell yet from the sample.

▷ Significance versus usefulness (See Figure 1).

From the definitions one can compute that $F = (n-2)\left(\frac{R^2}{1-R^2}\right)$. Form this we see that $R^2 = 0$ if $F = 0$ and vice versa. Also, $R^2 \to 1$ if $F \to \infty$ and vice versa. It appears as if R^2 and F are measuring the same thing. But usefulness and significance are different concepts. Figure 1 gives four extreme cases, illustrating the difference between the two statistics.

▷ Testing slope β_1 and intercept β_0.

Each of the values β_i can be tested for significance separately with a t-test. We have seen above that the F-test checks significance of the slope β_1. In general, (see next chapter) the F test checks significance of the slopes of all X-variables simultaneously. In simple regression there is only one X-variable and the F-test for significance is equivalent with the t-test for zero slope: the hypotheses are the same and so will be the conclusion, even though the test statistics (F and t) differ. (See the paragraph 'Identical tests'.)

$$\begin{aligned} H_0: &\quad \beta_1 = 0 \\ H_1: &\quad \beta_1 \neq 0. \end{aligned}$$

The test statistic is

$$t = \frac{b_1 - 0}{s_{b_1}} \sim t_{n-2}.$$

The standard error of the slope s_{b_1} can usually be read from the output. The assumptions for the test are those of the regression model. See 'checking the assumptions'.

Of course, we can also test one sided or for some non-zero slope, say δ:

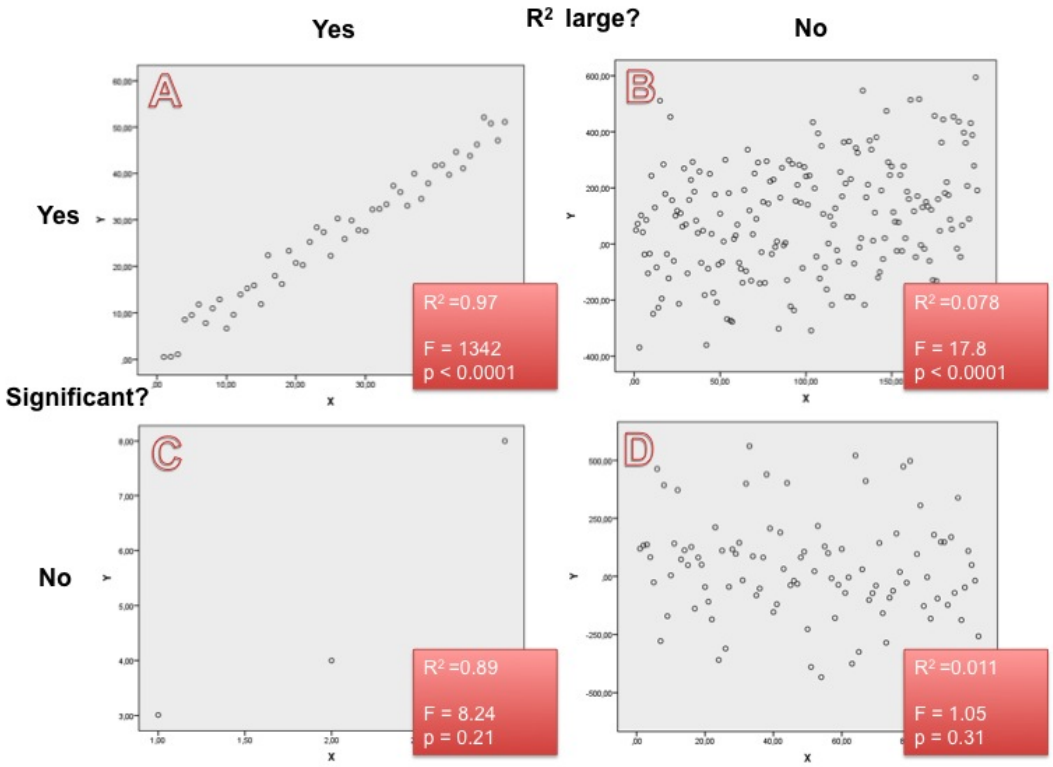

Figure 1: *Difference between R^2 ('usefulness') and F ('significance') in the regression analysis.* ***A****: Almost perfectly on the line.* ***B****: Clearly, there is a trend upwards. However, this slight shift upwards is small compared to the variance in Y. (Variance in Y only slightly explained by variation in X) The model is not useful in general but could be useful in some applications.* ***C****: Not enough points to conclude that the slope is significantly different from zero. We do not know yet yet wether the modle is useful, even though R^2 is close to 1.* ***D****: No trend up or down visible.*

$$\begin{array}{lll} H_0 : \beta_1 \geqslant \delta & & H_0 : \beta_1 = \delta & & H_0 : \beta_1 \leqslant \delta \\ H_1 : \beta_1 < \delta & \text{or} & H_1 : \beta_1 \neq \delta & \text{or} & H_1 : \beta_1 > \delta \end{array}$$

In the output, usually the t- and p-value's are only for the $H_0 : \beta_1 = 0$ and therefore these cannot be used directly for non-zero H_0. In that case, take the standard error of the slope s_{b_1} from the output and compute the t-value yourself:

$$t = \frac{b_1 - \delta}{s_{b_1}} \sim t_{n-2}.$$

Then, compare with the critical t from the table.

The t-test for the intercept β_0 works exactly the same as for β_1 but is usually less interesting.

Confidence intervals for β_0 and β_1:

$$\begin{aligned} b_0 - t_{n-2;\alpha/2} s_{b_0} &\leqslant \beta_0 \leqslant b_0 + t_{n-2;\alpha/2} s_{b_0}, \\ b_1 - t_{n-2;\alpha/2} s_{b_1} &\leqslant \beta_1 \leqslant b_1 + t_{n-2;\alpha/2} s_{b_1}. \end{aligned}$$

▷ Prediction / confidence intervals for Y

Given some observation with value $X = x$, what will be the corresponding value for Y? According to equation (16) we estimate it will be $\hat{y} = b_0 + b_1 x$. Of course, for one individual observation x there will be some error: It is unlikely that Y is exactly this value. But what if we have a very large number of observations, all with the same value $X = x$? Will the mean of the corresponding values for Y be $\hat{y} = b_0 + b_1 x$? If so, then that implies that we are spot on with our estimated regression equation. Of course, this is not very likely either. So both for individual predictions as well as for an estimate of the mean of the corresponding Y values, we can make an interval in which the values will be with some probaility $1 - \alpha$. (Of course, this only holds under the assumption of the regression model.) Note that the second interval is wider because individual Y-values vary more than their mean. In the formulas, $\hat{y} = b_0 + b_1 x$, $t = t_{n-2;\alpha/2}$, $s = \sqrt{\frac{SSE}{n-2}}$ (standard error), and $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

- **Confidence interval for the mean of Y.** Given value $X = x$, the value $\beta_0 + \beta_1 x$ will be in the interval with probability $1 - \alpha$:

$$\hat{y} \pm t \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}}.$$

- **Prediction interval for individual Y.** Given $X = x$, the corresponding value for Y will be in the interval with probability $1 - \alpha$:

$$\hat{y} \pm t \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}}.$$

▶ Identical tests.

In a simple regression, the following three tests will give the same conclusion.

- The t test for zero correlation. ($H_0 : \rho = 0$).
- The t test for zero slope. ($H_0 : \beta_1 = 0$)
- The F test for significance. ($H_0 : \beta_1 = 0$)

More precisely, the t-values of the first two are the same and the F-value of the last satisfies $F = t^2$. The p-values of the three test are identical so conclusions will be the same.

Chapter 13: Multiple Regression

▷ Regression model

In multiple regression we have more than one independent variables: $X_1, X_2, \ldots, X_k$. The Y-variable is the only dependent variable. Instead of a pair (x_i, y_i), each observation is given by $k+1$ values $x_{1i}, x_{2i}, \ldots, x_{ki}, y_i$.

- The *regression equation* becomes:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2),$$

  or with an index i for the i-th observation:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2).$$

- The *estimated regression equation* is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k,$$

  or with an index i for the i-th observation:

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}.$$

The *residual* for observation i is the difference between y_i and the estimate $\hat{y}_i$.

$$e_i = y_i - \hat{y}_i.$$

Just as in simple regression, the OLS method finds the estimates $b_0, b_1, \ldots, b_k$ which minimize $\sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2$.

▷ Assumptions

See simple regression.

▷ The computation (the ANOVA table).

The sums of squares are defined exactly the same as for simple regression. Only the degrees of freedom differ as can be seen from the mean sum of squares ($k = 1$ in simple regression):

$$MST = \frac{SST}{n-1}, \qquad MSR = \frac{SSR}{k}, \qquad MSE = \frac{SSE}{n-k-1}.$$

▷ Testing significance.

While in simple regression the F-test is doing the same as the t-test for β_1, in *multiple* regression it tests all the slopes $\beta_1, \beta_2, \ldots, \beta_k$ *simultaneously*. If rejected, then at least one of the slopes β_i differs significantly from zero. Then, a t-test for each of the β_i's can be used to check each of the coefficients seperately for non-zero slope.

$$H_0: \quad \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

$$H_1: \quad \beta_i \neq 0 \text{ for some } i \qquad \text{(or simply say: 'not } H_0\text{')}.$$

The test statistic is

$$F = \frac{MSR}{MSE} \sim F_{k,n-k-1} \qquad \text{(Reject for large values.)}$$

▷ Usefulness of the model.

See simple regression for a discussion on R^2 and the F-statistic. For multiple regression we also have the *adjusted coefficient of determination* R^2_{adj}.

$$R^2_{\text{adj}} = 1 - \left((1 - R^2) \frac{n-1}{n-k-1} \right).$$

If more variables are added to the model then R^2 increases. This is not necessarily true for R^2_{adj}. We always have $R^2_{\text{adj}} \leqslant R^2$. If R^2_{adj} is much less than R^2 then this indicates that the model contains useless predictors. We can remove one or more X-variables without loosing much of the predictive power.

▷ Testing slope β_i.

Same as for simple regression. The only adjustment is the degrees of freedom:

$$H_0 : \beta_i = \delta \quad \rightarrow \quad t = \frac{b_i - \delta}{s_{b_i}} \sim t_{n-k-1}.$$

▷ Dummy regressors.

A categorial variable X with only two possible values can be added to the regression model by recoding the two values into zero's and one's: $X = 0$ or $X = 1$.

▷ Detecting multicollinearity (Course: Statistics II).

When the independent variables $X_1, X_2, ..., X_k$ are intercorrelated instead of independent, we call this multicollinearity. (If only two predictors are correlated, we call this collinearity.) Multicollinearity is measured by

$$VIF_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the coefficient of determination of X_j with all other X-variables. If $VIF_j > 5$, then X_j is highly correlated with the other independent variables.

Chapter 15: Chi-Square Test

Two categorical variables A and B (or for example X and Y).

$$\begin{aligned} H_0: &\quad \text{Variables A and B are independent.} \\ H_1: &\quad \text{Variables A and B are dependent.} \end{aligned}$$

Equivalently:

$$\begin{aligned} H_0: &\quad \text{Row distributions are the same.} \\ H_1: &\quad \text{Row distributions are not the same.} \end{aligned}$$

or,

$$\begin{aligned} H_0: &\quad \text{Column distributions are the same.} \\ H_1: &\quad \text{Column distributions are not the same.} \end{aligned}$$

R_j: Total in row j $(j = 1, \ldots, r)$
C_k: Total in column k $(k = 1, \ldots, c)$.

Under H_0, the *expected* number of observations in cell j, k is

$$e_{jk} = n\left(\frac{R_j}{n}\right)\left(\frac{C_k}{n}\right) = \frac{R_j C_k}{n} \quad \text{(expected frequency)}.$$

Let f_{jk} be the *observed* number of observations in cell j, k. (Also denoted by n_{jk})

Test statistic:

$$\chi^2 = \sum_{j,k} \frac{(f_{jk} - e_{jk})^2}{e_{jk}} \sim \chi^2_{(r-1)(c-1)} \quad \text{(reject large values)}.$$

Condition: All expected frequencies e_{jk} should be at least 5. (SPSS uses: $e_{jk} \geqslant 5$ for at least 80% of the cells and $e_{jk} \geqslant 1$ for all cells.)

- If the condition is not met then combining rows or columns may solve this. The becomes less powerful though.
- The $2 \times c$ problem can be seen as comparing proportions: $H_0 : \pi_1 = \cdots = \pi_c$.
- For the 2×2 problem, the Chi-Square Test is equivalent with the z-test for equality of two proportions (See (15), Chapter 10). That means, the outcome (reject or not) will be the same.
- It is also possible to do a χ^2 test for numerical variables. In that case, each numerical variable is divided into categories. (For example, a numerical variable 'age' may be divided into 3 groups: less than 18, from 18 till 64, and at least 65.)

Chapter 16: Nonparametric Tests

Advantages: small samples allowed, few assumptions about population needed, ordinal data possible.

Disadvantages: less powerful if stronger assumptions can be made, requires new tables.

Two tests for the medians of paired samples:

- ▷ Wilcoxon Signed-Ranks Test
- ▷ Sign test (See Supplement 16A)

Two tests for the medians of independent samples:

- ▷ Wilcoxon Rank Sum test (also called the Mann-Whitney test). For two samples.
- ▷ Kruskal-Wallis test. For two or more samples .

One test for the correlation coefficient:

- ▷ Spearman rank correlation test.

▷ Wilcoxon Signed-Ranks Test

Test for the difference of medians in paired samples Y_1, Y_2 ($H_0 : M_1 = M_2$) or for the median in one sample Y ($H_0 : M_Y = M_0$)

Assumption: The population should be symmetric (for one sample) or the population difference should be symmetric (paired samples).

(Note: If a distribution is symmetric then its mean is the same as its median. Therefore, a test for the median can also be used as a test for the mean here.)

For paired samples, let variable $D = Y_1 - Y_2$ be the difference. (The book uses d.)
If we want to test one sample against a benchmark M_0 then we can take $D = Y - M_0$.
Using this variable D, the test is the same for both cases:

H_0: $M_D = 0$

Omit the zero's: $n \to n'$. Next, order from small to large by absolute values $|D_i|$. Add the signs to the ranks.

Sample statistic: W = sum of positive ranks.

Test statistic:

- For $n' < 20$ use tables and compare W with the critical values.
- For $n' \geqslant 20$ we use a normal approximation.

$$Z = \frac{W - \mu_W}{\sigma_W} \sim N(0,1), \text{ with } \mu_W = \frac{n'(n'+1)}{4}, \quad \sigma_W = \sqrt{\frac{n'(n'+1)(2n'+1)}{24}}.$$

NB. It is also possible to do a 1-sided test or even test against another value than 0. For example H_0: $M_D \geqslant -8$. (See for example Q3, tutorial 4) In that case, first subtract right side (-8) from D and give the variable a new name $D' = D - (-8)$. Now do the test $H_0 : M_{D'} \geqslant 0$.

▷ Sign test

The sign test is used for the same situations as the Wilcoxon Signed-Ranks test, i.e., for paired samples or for the median in one sample.

Assumptions: None.

Since symmetry is not required, we cannot test for μ but only for the median M. If symmetry can be assumed then Wilcoxon Signed-Ranks is preferred.

Again, let variable $D = Y_1 - Y_2$ be the differences of the two variables, or if want to test one sample Y against a benchmark M_0 then we can take $D = Y - M_0$. Omit the zero's: $n \to n'$.

H_0: $M_D = 0$

Sample statistic:

$$X = \text{ number of positive differences. Under } H_0 : X \sim \text{Bin}(n', \tfrac{1}{2}).$$

Test statistic:

- For $n' < 10$, use X and use the binomial distribution.
- For $n' \geqslant 10$, use normal approximation for the binomial (with continuity correction).

$$Z = \frac{X - \mu_x}{\sigma_x} \sim N(0,1), \quad \text{with} \quad \mu_X = \tfrac{1}{2}n', \text{ and } \sigma_X = \tfrac{1}{2}\sqrt{n'}.$$

NB. It is also possible to do a 1-sided test or even test against another value than 0. For example H_0: $M_D \geqslant -8$. (See for example Q4, tutorial 4) In that case, first subtract the right side (-8) from D and give the variable a new name $D' = D - (-8)$. Now do the test $H_0 : M_{D'} \geqslant 0$.

▷ Wilcoxon Rank Sum test (also called Mann-Whitney test).

Equality of medians of two independent samples.

Assumption: The two distributions have the same shape.

$H_0 : M_1 = M_2$.

Make combined rank (average rank for ties). Let T_1, T_2 be the sum of ranks in each sample, where the smallest sample is labeled 1.

Sample statistic: T_1

Test statistic:

- Small samples ($n_1 \leqslant 10$ and $n_2 \leqslant 10$): T_1. Reject both sides. Use table for critical values.
- Large samples ($n_1 \geqslant 10$ and $n_2 \geqslant 10$): Use the normal approximation. There are two equivalent versions. Either use

$$Z = \frac{T_1 - \mu_{T_1}}{\sigma_{T_1}}, \text{ with } \mu_{T_1} = \frac{n_1(n+1)}{2} \text{ and } \sigma_{T_1} = \sqrt{\frac{n_1 n_2 (n+1)}{12}}, \quad (n = n_1 + n_2),$$

  or use (see book):

$$Z = \frac{\bar{T}_1 - \bar{T}_2}{(n_1 + n_2)\sqrt{\frac{n_1 + n_2 + 1}{12 n_1 n_2}}} \sim N(0, 1).$$

Of course, we can also do a 1-sided test. If only the two-sided p-value is given then look at the mean ranks to decide on $p/2$ or $1 - p/2$. For example, if $H_0 : M_1 \leqslant M_2$ and $\bar{T}_1 > \bar{T}_2$ then our p-value should be less than 0.5. So we take $p/2$ and not $1 - p/2$.

Note that it is not defined what to do if $n_1 < 10$ and $n_2 > 10$. The table cannot be used since it only goes up to 10. The best we can do is to use the large sample test and mention that the condition $(n_1, n_2 \geqslant 10)$ is actually not satisfied.

▷ Kruskal-Wallis

Generalization of the Mann-Whitney test to more than two independent samples. It is therefore the non-parametric alternative for ANOVA (comparison of means).

Assumption: The c distributions have the same shape.

Condition: At least 5 observations in each sample.

$H_0 : M_1 = M_2 = \cdots = M_c$.
H_1 : not all medians are the same.

Order all observed values from smallest to largest and rank them $1, 2, \ldots, n$. (Average ranks for ties.)

Sample statistics: T_j: the sum of ranks of each group j.

Test statistic:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{c} \frac{T_j^2}{n_j} - 3(n+1) \quad \sim \quad \chi^2_{c-1} \quad \text{(reject large values)},$$

where n_j is the number of observations in group j and $n = n_1 + \cdots + n_c$.

▷ Spearman Rank correlation test.

This is a nonparametric test of association (correlation) between two variables. It is useful when it is inappropriate to assume an interval scale (a requirement of the Pearson correlation coefficient of Chapter 12).
We can test for ρ being unequal to zero:

$$\begin{array}{ll} H_0 : & \rho = 0 \quad \text{(No correlation)} \\ H_1 : & \rho \neq 0 \quad \text{(There is correlation)}. \end{array}$$

Replace every x_i by its rank within the X-values. Denote the new variable by X_r.
Replace every y_i by its rank within the Y-values. Denote the new variable by Y_r.

For these new variables, compute the correlation coefficient: $r_S = r(X_r, Y_r)$.
Test statistic:

$$Z = r_S \sqrt{n-1} \sim N(0,1).$$

Required: $n \geqslant 20$.