Tomás Pérez
Mathematical Biostatistics Boot Camp - Lecture Notes
July 29, 2022

Theory & Notes

1. Part 1

1.1. Week 1.

The expected value of the average of a collection of random variables from the same distribution is the same as the expected value of the individual random variables. Thus, the expected value of the sample mean is the population mean that it's trying to estimate.

1.2. Week 2.

Bayes' rule. Bayes' rule establishes a relationship between the conditional probabilities of A given B with the conditional probabilities of B given A.

Let $f(x|y)$ be the conditional density or mass function for $\mathbf{X}$ given that $\mathbf{Y} = y$. Let $f(y)$ be the marginal distribution for $\mathbf{Y}$:

$$\text{If } \mathbf{Y} \text{ is continuous then } f(y|x) = \frac{f(x|y)f(y)}{\int dt f(x|t)f(t)},$$

$$\text{If } \mathbf{Y} \text{ is discrete then } f(y|x) = \frac{f(x|y)f(y)}{\sum_t f(x|t)f(t)}.$$

A special case for this relationship is given for the case of two events A and B in the event space $\mathcal{F}(\Omega)$,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)},$$

which can be readily proved by setting $\mathbf{X}$ to be an indicator that event A has occurred and, similarly, letting $\mathbf{Y}$ be an indicator that event B has occurred and plugging said information into the discrete version of Bayes' rule.

Diagnostic testing. Consider a diagnostic test for some disease. Let + and - be the events in which the diagnostic test result came back as positive and negative respectively ie. the sample space is $\Omega = \{+, -\}$. Let D and D^c be the event in which the test subject does or doesn't have the disease respectively ie. the event space is $\mathcal{F} = \{D, D^c\}$. We can define two quantities of interest

- The **sensitivity** is the probability that the test is positive given that the subject actually has the disease, $\mathbb{P}(+|D)$.
- The **specificity** is the probability that the test is negative given that the subject doesn't have the disease, $\mathbb{P}(-|D^c)$.
- The **positive predictive value** is the probability that the subject has the disease given that the test is positive, $\mathbb{P}(D|+)$ (it quantifies the test's predictive "power").
- The **negative predictive value** is the probability that the subject doesn't have the disease given that the test is negative, $\mathbb{P}(D^c|-)$.
- We define the **prevalence of the disease** as the marginal probability of disease, $\mathbb{P}(D)$.

We define the **diagnostic likelihood ratio of a positive test**, labelled DLR_+, as

$$DLR_+ = \frac{\mathbb{P}(+|D)}{\mathbb{P}(+|D^c)} = \frac{\text{sensitivity}}{1 - \text{specificity}},$$

ie. the probability the test subject has the disease given the test result was positive divided by the probability the test subject doesn't have the disease given the test result was positive, whilst **diagnostic likelihood ratio of a negative test**, labelled DLR_-, is

$$DLR_- = \frac{\mathbb{P}(-|D)}{\mathbb{P}(-|D^c)} = \frac{1 - \text{sensitivity}}{\text{specificty}}.$$

ie. probability of a negative test result given the test subject actually has the disease divided by the probability of a negative test result given the subject doesn't have the disease.

Consider now the following example:

Diagnostic test example

A study comparing the efficacy of HIV tests, reports an experiment which concluded that HIV antibody tests have a sensitivity of 99.7% and a specificity of 98.5%. Suppose that a subject, from a population with a .1% prevalence of HIV, receives a positive result. What is the probability that the subject has HIV?

We desire to compute $\mathbb{P}(D|+)$ given the sensitivity is $\mathbb{P}(+|D) = .997$, the specificity is $\mathbb{P}(-|D^c) = .985$ and the prevalence being $\mathbb{P}(D) = .001$. Thus, according to Bayes' rule

$$\begin{aligned}
\mathbb{P}(D|+) &= \frac{\mathbb{P}(+|D)\mathbb{P}(D)}{\mathbb{P}(+|D)\mathbb{P}(D) + \mathbb{P}(+|D^c)\mathbb{P}(D^c)} \\
&= \frac{\mathbb{P}(+|D)\mathbb{P}(D)}{\mathbb{P}(+|D)\mathbb{P}(D) + (1 - \mathbb{P}(-|D^c))(1 - \mathbb{P}(D))} \text{ since } \mathbb{P}(+|D^c) = 1 - \mathbb{P}(-|D^c) \\
&= \frac{.997 \times .001}{.997 \times .001 + .015 \times .999} = .062,
\end{aligned}$$

Thus, in this population, a positive test result only suggests a 6% probability that the subject has the disease. The positive predictive value for this test is 6%.

Now, note that the test's sensitivity (has the disease given a positive test result) is 99.9% for any arbitrary test subject whilst the specificity (doesn't have the disease given a negative test result) is 98.5%. Why is the positive predictive value so low? this is due to low prevalence of disease and the somewhat modest specificity. This is entirely due to Bayes' rule's nature: we start with a very low probability of thinking that this person has the disease, it's updated with the information of the positive test result by taking into account the test's sensitivity and specificity and informs the positive predictive value. In this particular case, we started with such a low prior.
In contrast, suppose it was known that some test subject was an intravenous drug user and routinely had intercourse with an HIV infected user, now the prior is much higher than the low prevalence of the previous example. Thus we'd expect a much higher positive predictive value. Notice that the evidence implied by a positive test result doesn't change because of the prevalence of the disease in the subject's population, only our interpretation of that evidence changes.

The previous example begs the question of what's the component of the calculation which doesn't depend on changes on the prevalence? the diagnostic likelihood ratios. In effect, according to Bayes' rule

$$\begin{aligned}
\mathbb{P}(D|+) = \frac{\mathbb{P}(+|D)\mathbb{P}(D)}{\mathbb{P}(+|D)\mathbb{P}(D) + \mathbb{P}(+|D^c)\mathbb{P}(D^c)} &\text{ and } \mathbb{P}(D^c|+) = \frac{\mathbb{P}(+|D^c)\mathbb{P}(D^c)}{\mathbb{P}(+|D)\mathbb{P}(D) + \mathbb{P}(+|D^c)\mathbb{P}(D^c)} \\
\rightarrow \text{therefore } \frac{\mathbb{P}(D|+)}{\mathbb{P}(D^c|+)} &= \frac{\mathbb{P}(+|D)}{\mathbb{P}(+|D^c)} \times \frac{\mathbb{P}(D)}{\mathbb{P}(D^c)}, \\
\rightarrow \text{post-test odds of } D &= DLR_+ \times \text{pre-test odds of D}
\end{aligned}$$

thus the DLR_+ is a multiplicative factor which dampens or amplifies the probabilities based on the pre-test odds.

Diagnostic test example continued

Continuing from our previous example, suppose a subject has a positive HIV test. Thus, the $DLR_+ = \frac{.997}{1-.985} = 66$, then the result of the positive test is that the odds of disease is now 66 times the pretest odds. Or equivalently, the hypothesis of disease is 66 times more supported by the data than the data of no disease.

Now, suppose a subject has a negative test result, then the diagnostic likelihood ratio for a negative test is $DLR_- = \frac{1-.997}{.985} = .003$. Therefore the post-test odds of disease is now .3% of the pretest odds given the negative test. Or equivalently, the hypothesis of disease is supported .003 times than that of the hypothesis of absence of disease given the negative test result[a].

[a] According to Bayesian's statistics, Bayes' rule establishes the change of belief in theory given the data. The posterior belief is multiplied by the DLR vis-a-vis the data. According to the frequentist interpretation, a given subject either has the disease or they don't, leaving no room for probability. But it's interpreted in the infinite fictitious repetitions of this experiment.

Likelihood. A common and fruitful approach to statistics to statistics is to assume the data arises from a family of distributions indexed by a parameter which represents a useful summary of the distribution. The **likelihood** of a collection of data is the join density evaluated as a function of the parameters with the data fixed. Likelihood analysis of data uses the likelihood to perform inference regarding the unknown parameter. With this approach, the estimators can be used to know the estimands.

Given a statistical probability mass function or density, denoted by $f(\mathbf{x}, \theta)$ where $\mathbf{x} \in \mathbb{R}^n$ is a random vector and where $\theta \in \Theta$ is an unknown parameter in the parameter space Θ. Then the likelihood is f viewed as function of θ for a fixed, observed value of $\mathbf{x}$.

The likelihood has the following properties:

- Ratios of likelihood values measure the relative evidence of one value of the unknown parameter to another. This is called the **law of likelihood**, which is the notion that the extent to which the evidence supports one parameter value or hypothesis against another is indicated by the ratio of their likelihood ie.

$$\Lambda = \frac{\mathcal{L}(a|\mathbf{X}=x)}{\mathcal{L}(b|\mathbf{X}=x)} = \frac{\mathbb{P}(\mathbf{X}=x|a)}{\mathbb{P}(\mathbf{X}=x|b)}.$$

  This number is the degree to which the observation x supports parameter value or hypothesis a against b. If this ratios is 1, evidence is indifferent, if greater than 1 the evidence supports a against b, then vice versa- In Bayesian statistics, this ratio is the Bayes' factor. This is a special case of the Neymann-Pearson lemma.

- Given a statistical model and observed data, all of the relevant information contained in the data regarding the unknown parameter is contained in the likelihood. This theorem is called the **likelihood principle**.

- If $\{\mathbf{X}_i\}_{i=1}^m$ are independent random variables, then the global likelihood is the product of the individual likelihoods. This is, the likelihood of the parameters given all of the $\mathbf{X}_i$ is simply the product of thee individual likelihoods.

Consider now the following example,

Likelihoods example

Consider a coin-flip experiment, it's results being described by a dichotomic Bernoulli random variable $\mathbf{X} \sim \text{Bernoulli}(1, \theta)_{\theta \in \Theta \simeq \mathbb{R}_{[0,1]}}$. The mass function for x is

$$f(x, \theta) = \theta^x (1-\theta)^{1-x} \text{ for } \theta \in \mathbb{R}_{[0,1]},$$

where x is either 0 for tails or 1 for heads. Suppose the result is a head, the likelihood is

$$\mathcal{L}(\theta, 1) = \theta^1 (1-\theta)^{1-1} = \theta.$$

Therefore, for example, $\frac{\mathcal{L}(.5,1)}{\mathcal{L}(.25,1)} = \frac{.5}{.25} = 2$, there is twice as much evidence supporting the hypothesis of $\theta = .5$ than the hypothesis of $\theta = .25$.

Consider now multiple independent coin flips, in this case, four, obtaining the sequence $1, 0, 1, 1$. Thus, we're dealing with four Bernoulli random variables $\{\mathbf{X}_i | \mathbf{X}_i \sim \text{Bernoulli}(1, p)\}_{i=1}^4$, note that the sum of said random variables is a binomial random variable: $\mathbf{Y} = \sum_{i=1}^4 \mathbf{X}_i \sim \text{Binom}(n = 4, p)$. The likelihood is obtained as the product of the four, independent, probability mass functions

$$\mathcal{L}(\theta, a_1 = 1, a_2 = 0, a_3 = 1, a_4 = 1) = \prod_{i=1}^4 \theta^{a_i} (1-\theta)^{a_i} = \theta^3 (1-\theta)^1.$$

This likelihood only depends on the total number of heads and the total number of tails, as such we can denote it $\mathcal{L}(\theta, 1, 3)$. For example, given that $\frac{\mathcal{L}(.5,1,3)}{\mathcal{L}(.25,1,3)} = 5.33$, there is over five times as much evidence supporting the hypothesis of $\theta = .5$ over than $\theta = .25$.

Refering to the previous example, in general, we're not interested in performing pair-wise comparisons of the likelihoods for different parameter values. Generally, we want to consider all values $\theta \in \mathbb{R}_{[0,1]}$. A **likelihood plot** displays $\mathcal{L}(\theta, x)$ vs $.\theta$, normalized so its height is 1.

The value of θ where the curve reaches its maximum is the data's most well supported value of θ. Said point is called the **maximum likelihood estimate** of θ

$$\hat{\theta} = \underset{\theta \in \Theta}{\text{argmax}}\, \mathcal{L}(\theta, x).$$

Another interpretation for the MLE is that it's the value of θ that would make the data we observed the most probable.

Some technical aspects. **An introduction to Measure Theory**

A **measure** is a mathematical device which reflects the notion of quantity for a given set. Let $\mathbf{X}$ be a set, then each subset $\mathbf{U} \in \mathbf{X}$ is assigned a positive real number $\mu[\mathbf{U}]$. Thus, the measure is a function

$$\mu : \mathcal{P}(\mathbf{X}) \to \mathbb{R},$$

where $\mathcal{P}(\mathbf{X}) = \{\mathbf{S} \in \mathbf{X}\}$ is the power set of $\mathbf{X}$. However, it's usually impossible to define a satisfactory notion of quantity for all subsets of $\mathbf{X}$. Therefore, $\text{Dom}(\mu) \in \mathbf{X}$ ie. only some subsets of $\mathbf{X}$ will be measurable. Before defining a proper measure, it's domain must be specified first. This domain will be a collection of subsets of the space $\mathbf{X}$, called a **sigma algebra**.

Let $\mathbf{X}$ be a set. A **σ-algebra** over is a collection $\mathcal{F}$ of subsets of $\mathbf{X}$ with the following properties:

- $\mathcal{F}$ is closed under countable unions ie.

$$\{\mathbf{U}_i\}_{i\in\mathbb{N}} \mid \mathbf{U}_i \subset \mathcal{F} \Rightarrow \underset{i\in N}{\cup}\mathbf{U}_i \subset \mathcal{F}.$$

- $\mathcal{F}$ is closed under countable intersections ie.

$$\{\mathbf{U}_i\}_{i\in\mathbb{N}} \mid \mathbf{U}_i \subset \mathcal{F} \Rightarrow \underset{i\in N}{\cap}\mathbf{U}_i \subset \mathcal{F}.$$

- $\mathcal{F}$ is closed under complementation: if $\mathbf{U} \subset \mathcal{F}$, then

$$\mathbf{U}^c = \mathbf{X}/\mathbf{U} \subset \mathcal{F}.$$

A **measurable space** is an ordered pair $(\mathbf{X}, \mathcal{F})$, where $\mathbf{X}$ is a set and where $\mathcal{F}$ is a sigma-algebra on $\mathbf{X}$. Some of the most common examples of σ-algebras are

- Two examples of trivial σ-algebras are:
  - for any set $\mathbf{X}$ the collection $\{\emptyset, \mathbf{X}\}$ is a σ-algebra. This first example is far too small to be of any use.
  - The power set $\mathcal{P}(\mathbf{X})$ is also a σ-algebra. Note that for large sets, the power set becomes too large to be manageable.

- A more manageable σ-algebra is the one induced by the (co-)countable sets. Let $\mathcal{M}$ be the most conservative collection of ”manageable” sets, this is

$$\mathcal{M} = \Big\{\{x\} : x \in \mathbf{X}\Big\},$$

  ie. the set of all the singleton subsets of $\mathbf{X}$. Then $\mathcal{C} = \sigma(\mathcal{M}))$ is the σ-algebra of **countable and co-countable sets**

$$\mathcal{C} = \{\mathbf{C} \subset \mathbf{X} | \text{ either } \mathbf{C} \text{ is countable, or } \mathbf{X}/\mathbf{C} \text{ is countable}\}.$$

  If $\mathbf{X}$ is itself finite or countable, then $\mathcal{C} = \mathcal{P}(\mathbf{X})$.

- Another example are the partition algebras. Let $\mathbf{X}$ be a set. Then a **partition** of $\mathbf{X}$ is a collection $\mathcal{P} = \{\mathbf{P}_i\}_{i=1}^N$ of disjoint subsets, such that $\mathbf{X} = \sqcup_{n=1}^N \mathbf{P}_N$. These subsets $\mathbf{P}_i$ are called the **atoms** of the partition. Then, the σ-algebra generated by $\mathcal{P}$ is the collection of all possible unions of $\mathcal{P}$-atoms:

$$\sigma(\mathcal{P}) = \{\sqcup_{j=1}^k \mathbf{P}_{n_j} \mid \{n_j\}_{j=1}^k \in \mathbb{N}_{[1,\cdots,N]}\}.$$

  Therefore if card$[\mathbf{P}] = N$, then card$[\sigma(\mathbf{P})] = 2^N$.

  If $\mathcal{Q}$ is another partition, we say $\mathcal{Q}$ redefines $\mathcal{P}$ ($\mathcal{Q} \prec \mathcal{P}$) if, for every $\mathbf{P} \in \mathcal{Q}$, there are $\{\mathbf{Q}_i\}_{i=1}^N \in \mathcal{Q}$ so that $\mathbf{P} = \sqcup_{j=1}^N \mathbf{Q}_j$. In said case we have

$$\mathcal{P} \prec \mathcal{Q} \Leftrightarrow \sigma(\mathcal{P}) \subset \sigma(\mathcal{Q})$$

.

- **Borel σ-algebra of** $\mathbb{R}$:

  Let $\mathbf{X} = \mathbb{R}$ be the real numbers and let $\mathcal{M}$ be the set of all open intervals in $\mathbb{R}$:

  $$\mathcal{M} = \{(a,b) : -\infty \leq a < b \leq \infty\}$$

  ,
  then the σ-algebra $\mathcal{B} = \sigma(\mathcal{M}$ contains all open subsets of $\mathbb{R}$, all closed subset, all countable intersections of open subsets, countable unions of closed subsets, etc. For example, $\mathcal{B}$ contains, as elements, the set $\mathbb{Z}$ of integers, the set $\mathbb{Q}$ of rationals and the set $\mathbb{I}$ of irrationals. Then, $\mathcal{B}$ is called the **Borel σ-algebra** of $\mathbb{R}$.

In general, let $\mathbf{X}$ be a topological space and let $\mathcal{M}$ be the set of all open subsets of $\mathbf{X}$. The σ-algebra $\sigma(\mathcal{M})$ is the **Borel σ-algebra** of $\mathbf{X}$ and is denoted by $\mathcal{B}(\mathbf{X})$. It contains all open sets and closed subsets of $\mathbf{X}$, all countable intersections of open sets (called $G\delta$ sets), all countable unions of closed sets (called $F\sigma$ sets). For example, if $\mathbf{X}$ is Hausdorff, then $\mathcal{B}(\mathbf{X})$ contains all countable and co-countable sets.

Let $(\mathbf{X}, \mathcal{F})$ be a measurable space. A **measure** on $\mathcal{F}$ is a map $\mu : \mathcal{F} \to \mathbb{R}_+$, which is **countably additive** ie.

$$\text{If } \{\mathbf{Y}_i\}_{i=1} \mid \mathbf{Y}_i \in \mathcal{F} \to \mu\Big[\sqcup_{n=1}^{\infty} \mathbf{Y}_n\Big] = \sum_{n=1}^{\infty} \mu[\mathbf{Y}_n].$$

Then a **measure space** is an ordered triple $(\mathbf{X}, \mathcal{F}, \mu)$, where $\mathbf{X}$ is a set, $\mathcal{F}$ is a σ-algebra and μ is a measure on $\mathcal{F}$. Thus, μ assigns a size to the $\mathcal{F}$-measurable subsets of $\mathbf{X}$. Some important measures and measureable spaces are

- **The counting measure** assigns, to any set, the cardinality of that set.

  $$\mu[\mathbf{S}] = \text{card}[\mathbf{S}].$$

  This measure provides no means of distinguishing between sets of the same cardinality, only being useful in finite measure spaces.

- A **finite measure space** is made up by a finite set $\mathbf{X}$, and a σ-algebra $\mathcal{F} = \mathbf{X}$. Then a measure μ on $\mathbf{X}$ is entirely defined by some function $f : \mathbf{X} \to \mathbb{R}_+$.For any subset $\{x_i\}_{i=1}^N$ we then define

  $$\mu\Big(\{x_i\}_{i=1}^N\Big) = \sum_{i=1}^N f(x_i).$$

- **Discrete measures**: If $(\mathbf{X}, \mathcal{F}, \mu)$ is a measure space, then an **atom** of μ is a subset $\mathbf{A} \in \mathcal{F}$ such that
  - $\mu[\mathbf{A}] = A > 0$.
  - For any $\mathbf{B} \subset \mathbf{A}$, either $\mu[\mathbf{B}] = A$ or $\mu[\mathbf{B}] = 0$.

  For example, in the finite measure space above, the singleton set $\{x_n\}$ is an atom if $f(x_n) > 0$. The measure space $(\mathbf{X}, \mathcal{F}, \mu)$ is called discrete if we can write

  $$\mathbf{X} = \mathbf{Z} \sqcup (\sqcup_{n=1}^{\infty} A_n),$$

  where $\mu[\mathbf{Z}] = 0$ and where $\{A_n\}_{n=1}^{\infty}$ is a collection of atoms. Note that any finite measure space is discrete.

- **The Lebesgue measure**, the **Haar measures** and the **Hausdorff measure**.

- **Stieltjes measures**: if we see $\mathbb{R}$ as a group, then the Lebesgue measure arises as a Haar measure, if we see $\mathbb{R}$ as a metric space, the Lebesgue measure arises as a Hausdorff measure. Instead, if we treat $\mathbb{R}$ as an ordered set, then the Lebesgue measure arises froma Stieltjes measure.

  Given an ordered set $(\mathbf{X}, <)$, we can define our σ-algebra $\mathcal{F}$ to be the σ-algebra generated by all left-open intervals of the form $(a, b]$[1]. Now suppose that $f : \mathbf{X} \to \mathbb{R}$ is a right-continuous, non-decreasing function, we can define the measure of any interval $(a, b]$ to be simply the difference between the value of f at the two end-points, a and b :

  $$\mu_f(a, b] = f(b) - f(a).$$

  We then extend this measure to the rest of the elements of $\mathcal{F}$ by approximating them with disjoint union of left-open intervals.

  We call μ_f a **Stieltjes measure** and call f the **accumulation function** or **cumulative distribution** of μ_f. Under suitable conditions, every measure on $(\mathbf{X}, \mathcal{F})$ can be generated in this way: Starting with an arbitrary measure, μ, we can find a zero-point $x_0 \in \mathbf{X}$ so that
  - $\mu(x_0, x]$ is finite for all $x > x_0$,

  - $\mu(x_0, x]$ is finite for all $x < x_0$

  - Then define the function $f : \mathbf{X} \to \mathbb{R}$ as

    $$f(x) = \begin{cases} \mu(x_0, x] & \text{if } x > x_0 \\ -\mu(x_0, x] & \text{if } x < x_0 \end{cases}$$

  .

- **Density Functions**: Let $\rho : \mathbb{R}^n \to \mathbb{R}$ be a positive, integrable function on $\mathbb{R}^n$. For any $\mathbf{B} \in \mathcal{B}(\mathbb{R}^n)$, we can define

  $$\mu_\rho(\mathbf{B}) = \int_{\mathbf{B}} \rho.$$

  We call ρ the **density function** for μ.

- **Probability Measures**: A measure μ on $\mathbf{X}$ is a **probability measure** if $\mu[\mathbf{X}] = 1$. Then the ordered triple $(\mathbf{X}, \mathcal{F}, \mu)$ is a **probability space**.

- **Stochastic processes** are a particular king of probability measures, which represent a system randomly evolving in time. Let $\mathcal{S}$ be any randomly-evolving complex system, let $\mathbf{X}$ be the set of all possible states of the system $\mathcal{S}$ and let $\mathbb{T}$ be a set representing time. For example,
  - If $\mathcal{S}$ is a rolling die, then $\mathbf{X} = \{1, 2, 3, 4, 5, 6\}$ and $\mathbb{T} = \mathbb{N}$ indexes the successive dice rolls.

  - If $\mathcal{S}$ is a publically traded stock, then its state is its price. Thus $\mathbf{X} = \mathbb{R}$. If we assume trading occurs continuously when the market is open, and let each trading day have length $c < 1$, then one representation of market time is $\mathbb{T} = \sqcup_{n=1}^{\infty} [n, n + c]$.

  - If $\mathcal{S}$ is a weather system, then its state can be representated by a large array of data $\mathbf{x} = [x_1, \cdots, x_n]$. Thus $\mathbf{X} = \mathbb{R}^n$ and since the weather evolves continuously $\mathbb{T} = \mathbb{R}$.

  We represent the random evolution of $\mathcal{S}$ by assigning a probability to every possible *history.* A history is an assignment of a state in $\mathbf{X}$ to every moment in time (ie. in $\mathbb{T}$). In other words, it's a

[1] If $\mathbf{X} = \mathbb{R}$ with the usual linear ordering, then this $\mathcal{F}$ is just the usual Borel σ-algebra.

function $f : \mathbb{T} \to \mathbf{X}$. The set of all possibly histories is $\mathbf{H} = \mathbf{X}^{\mathbb{T}}$ ie. the set of all functions $f : \mathbb{T} \to \mathbf{X}$.

The σ-algebra on $\mathbf{H}$ is usually a **cylinder algebra**.

– Let $(\mathbf{X}_\lambda, \mathcal{X}_\lambda)$ be measurable spaces for all $\lambda \in \Lambda$, where Λ is some (possibly uncountably infinite) indexing set. Consider the cartesian product $\bigtimes_{\lambda \in \Lambda} \mathbf{X}_\lambda$. Let

$$\mathcal{M} = \left\{ \bigtimes_{\lambda \in \Lambda} \mathbf{U}_\lambda \mid \forall \lambda \in \Lambda \text{ and } \mathbf{U}_\lambda \in \mathcal{F}_\lambda \text{ and } \mathbf{U}_\lambda = \mathbf{X}_\lambda \text{ for all but finitely many } \lambda \right\},$$

such subsets are called **cylinder sets** in $\mathbf{X}$ and $\sigma(\mathcal{M})$ is the **cylinder σ-algebra** denoted by $\bigtimes_{\lambda \in \Lambda} \mathbf{X}_\lambda \mathcal{F}_\lambda$. If the $\mathbf{X}_\lambda$ are topological spaces with Borel σ-algebras $\mathcal{F}_\lambda$, and we endow $\mathbf{X}$ with the Tychonoff product topology, then $\bigtimes_{\lambda \in \Lambda} \mathcal{F}_\lambda$ is the Borel σ-algebra of $\mathbf{X}$.

Suppose that $\mathbf{X}$ has a σ-algebra $\mathcal{F}$, then it follows $\mathbf{H}$'s σ-algebra is $\mathcal{H} = \bigtimes_{t \in \mathbb{T}} \mathcal{F}_t$. An **event** is an element of $\mathcal{F}$ and thus corresponds to a cylinder set, a countable union of cylinder sets etc.

Suppose, for all $t \in \mathbb{T}$, that $\mathbf{U}_t \in \mathcal{F}$ with $\mathbf{U}_t = X$ for all but finitely many t. The cylinder set $\mathbf{U} = \prod_{t \in \mathbb{T}} \mathbf{U}_t$ thus corresponds to the assertion: "for every $t \in \mathbb{T}$, at time t, the state of $\mathcal{S}$ was inside $\mathbf{U}_t$. A probability measure on $(\mathbf{H}, \mathcal{H})$ is then a way of assigning probabilities to such assertions.

Convergence of random variables:

Let $\{\mathbf{X}_n\}_{n=1}$ be a sequence of random variables, let $\mathbf{X}$ be a random variable, all of them defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. There are several, important, types of stochastic convergence, which will be treated as follows

- A sequence $\{\mathbf{X}_n\}_{n=1}$ of real-valued random variables is said to **converge in distribution** or to **weakly-converge** to a random variable $\mathbf{X}$ if

$$\mathbf{X}_n \overset{D}{\to} \mathbf{X} \text{ if } \mathbb{P}(\mathbf{X}_n \leq x) \underset{n \to \infty}{\to} \mathbb{P}(\mathbf{X} \leq x),$$

for every real number x at which the cumulative distribution functions are continuous. For random vectors, the definition is completely analogous. We say that a sequence of random vectors $\{\mathbf{X}_n\}_{n=1}^{\cdots} \subset \mathbb{R}^k$ converges in distribution to a random k-vector $\mathbf{X}$ if

$$\mathbf{X}_n \overset{D}{\to} \mathbf{X} \text{ if } \mathbb{P}(\mathbf{X}_n \in \mathbf{A}) \underset{n \to \infty}{\to} \mathbb{P}(\mathbf{X} \in \mathbf{A}),$$

for every $\mathbf{A} \in \mathbb{R}^k$ which is a continuity set of $\mathbf{X}$, ie. the set in which the function is continuous. Some important consequences of weak convergence are

  – $\mathbb{E}f(\mathbf{X}_n) \to \mathbb{E}f(\mathbf{X})\ \forall f \in C^0(\mathbb{R})$.

  – **Continous mapping theorem**: Let $g \in C^0(\mathbb{R})$, if the sequence $\{\mathbf{X}\}_{n=1}^{\cdots}$ converges in distribution to $\mathbf{X}$, then $\{g(\mathbf{X})\}_{n=1}^{\cdots}$ converges to $g(\mathbf{X})$.

- **Convergence in probability**: A sequence $\{\mathbf{X}_n\}_{n=1}^{\infty}$ of random variables converges to the random variable $\mathbf{X}$ if

$$\forall \epsilon \in \mathbb{R}_+ \ , \mathbb{P}(|\mathbf{X}_n - \mathbf{X}|) > \epsilon) \underset{n \to \infty}{\to} 0,$$

ie. let $\mathbb{P}_n(\epsilon)$ be the probability that $\mathbf{X}_n$ is outside the ϵ-ball centered at $\mathbf{X}$. Then $\{\mathbf{X}_n\}_{n=1}^{\infty}$ is said to converge in probability to $\mathbf{X}$ if, for any $\epsilon, \delta \in \mathbb{R}_+$, there exists a number $N(\epsilon, \delta)$ such that for all $n \geq N$, $\mathbb{P}_n(\epsilon) < \delta$.

$$\{\mathbf{X}_n\}_{n=1}^{\infty} \overset{p}{\to} \mathbf{X} \text{ if } \forall \epsilon,\ \delta \in \mathbb{R}_+, \exists N(\epsilon, \delta) \in \mathbb{R} \text{ such that } \forall n \geq N, (\mathbb{P}|\mathbf{X} - \mathbf{X}_n| - 0) < \delta.$$

Note that, for any sequence of random variables to converge in probability to another random variable, it isn't possible for $\mathbf{X}$ and $\mathbf{X}_n$ to be independent random variables, for each n. Thus convergence in probability is a condition on the joint cumulative distribution function, as opposed to convergence in distribution, which is a condition on the individual CDF's. For random elements $\{\mathbf{X}_n\}_{n=1}^{\infty}$ on a separable metric space (S, d), convergence in probability is similarly defined by

$$\forall \epsilon \in \mathbb{R}_+, \mathbb{P}(d(\mathbf{X}_n, \mathbf{X}) \geq \epsilon) \to 0.$$

Some important consequences of convergence in probability are:

– convergence in probability implies convergence in distribution:

$$\{\mathbf{X}_n\}_{n=1}^{\infty} \xrightarrow{p} \mathbf{X} \Rightarrow \{\mathbf{X}_n\}_{n=1}^{\infty} \xrightarrow{d} \mathbf{X}.$$

– Convergence in distribution implies convergence in probability only when the limiting random variable is a constant.

– The continuous mapping theorem, which states that

$$\forall g \in C^0(\mathbb{R}), \text{ if } \{\mathbf{X}_n\}_{n=1}^{\infty} \xrightarrow{p} \mathbf{X} \Rightarrow g(\{\mathbf{X}_n\}_{n=1}^{\infty}) \xrightarrow{p} g(\mathbf{X})$$

– Convergence in probability defines a topology on the space of random variables over a fixed probability space. This topology is metrizable by the Ky Fan-metric:

$$d(\mathbf{X}_a, \mathbf{X}_b) = \inf\{\epsilon \in \mathbb{R}_+ : \mathbb{P}(|\mathbf{X}_a - \mathbf{X}_b| > \epsilon) \leq \epsilon\},$$

or alternately by

$$d(\mathbf{X}_a, \mathbf{X}_b) = \mathbb{E}[\min(|\mathbf{X}_a - \mathbf{X}_b|, 1)].$$

- **Almost sure convergence**: We say a sequence $\{\mathbf{X}_n\}_{n=1}^{\infty}$ converges almost surely or almost everywhere or strongly converges towards $\mathbf{X}$ if

$$\mathbb{P}\Big(\lim_{n\to\infty} \mathbf{X}_n = \mathbf{X}\Big) = 1.$$

This means that the values of $\{\mathbf{X}_n\}_{n=1}^{\infty}$ approach the value of $\mathbf{X}$ in the sense that events for which $\{\mathbf{X}_n\}_{n=1}^{\infty}$ doesn't converge to $\mathbf{X}$ have probability 0. The previous statement is equivalent to

$$\{\mathbf{X}_n\}_{n=1}^{\infty} \xrightarrow{a.e.} \mathbf{X}, \text{ if } \mathbb{P}\Big(\omega \in \Omega : \mathbf{X}_n(\omega) \underset{n\to\infty}{\to} \mathbf{X}(\omega)\Big),$$

which in turn is equivalent to the following statement:

$$\{\mathbf{X}_n\}_{n=1}^{\infty} \xrightarrow{a.e.} \mathbf{X}, \text{ if } \mathbb{P}\Big(\limsup_{n\to\infty} \{\omega \in \Omega : |\mathbf{X}_n(\omega) - \mathbf{X}(\omega)| > \epsilon\}\Big) = 0, \forall \epsilon \in \mathbb{R}_+.$$

For generic random elements $\{\mathbf{X}_n\}_{n=1}^{\infty}$ on a metric space (S, d), strong converge is defined as

$$\mathbb{P}\Big(\omega \in \Omega : d(\mathbf{X}_n(\omega), \mathbf{X}(\omega)) \underset{n\to\infty}{\to} 0\Big) = 1.$$

Some consequences of strong convergence are

– By Fatou's lemma, almost sure convergence implies convergence in probability and hence implies convergence in distribution. This is the notion which justifies the strong law of large numbers.

- **Pointwise convergence** We say a sequence of random variables $\{\mathbf{X}_n\}_{n=1}^{\infty}$ pointwise converges towards $\mathbf{X}$ if

$$\lim_{n\to\infty} \mathbf{X}_n(\omega) = \mathbf{X}(\omega),\, \forall \omega \in \Omega,$$

in other words

$$\Big\{\omega \in \Omega | n \to \infty \mathbf{X}_n(\omega) = \mathbf{X}(\omega)\Big\} = \Omega.$$

- **Convergence in mean**: Given a real number $r > 1$, we say a sequence $\{\mathbf{X}_n\}_{n=1}^{\infty}$ converges in r-th mean (or in the L^r-norm) towards the $\mathbf{X}$-random variable if the r-th absolute moments, $\mathbb{E}(|\mathbf{X}_n|^r)$ and $\mathbb{E}(|\mathbf{X}|^r)$, of both $\{\mathbf{X}_n\}_{n=1}^{\infty}$ and $\mathbf{X}$ exist and

$$\{\mathbf{X}_n\}_{n=1}^{\infty} \xrightarrow{L^r} \mathbf{X} \text{ if } \lim_{n\to\infty} \mathbb{E}(|\mathbf{X}_n - \mathbf{X}|^r) = 0$$

.

Some important cases and consequences of r-th mean convergence are:

– When $\{\mathbf{X}_n\}_{n=1}^{\infty}$ converges in 1-th mean to $\mathbf{X}$, we say it converges in mean to $\mathbf{X}$.

– When $\{\mathbf{X}_n\}_{n=1}^{\infty}$ converges in 2-th mean to $\mathbf{X}$, we say it converges in mean square to $\mathbf{X}$.

– Note that r-th mean, for $r \geq 1$, implies convergence in probability by Markov's inequality:

$$\text{If } \mathbf{X} \text{ is a non-negative random variable and } a > 0, \text{then } \mathbb{P}(\mathbf{X} \geq a) \leq \frac{\mathbb{E}(\mathbf{X})}{a},$$

ie. the probability that X is at least a is at most the expectation of X divided by a: or in measure-theoretic language, let f be a measurable extended real-value function and $\epsilon \in \mathbb{R}_+$

$$\mu(\{x \in X : |f(x)| \geq \varepsilon\}) \leq \frac{1}{\varepsilon} \int_X |f| \, d\mu..$$

Law of large numbers:

The law of large numbers is a theorem that describes the result of performing the same experiment a large number of times. According to the law, the average of the results obtained from a large number of trials should be close to the expected value and tends to become closer to the expected value as more trials are performed. There are two versions of this theorem:

- The **Weak law of large numbers** states that the sample average converges in probability towards the expected value

$$\bar{\mathbf{X}} \xrightarrow{p} \mu \text{ that is } \lim_{n\to\infty} \mathbb{P}\Big(|\bar{\mathbf{X}} - \mu| < \epsilon\Big) = 1.$$

The weak law states that for any non-zero margin specified ϵ, no matter how small, with a sufficiently large sample there will be a very high probability that the average of the observations will be close to the expected value, that is, within the margin. This weak law applies in the case of iid random variables.

- The **Strong law of large number** states that the sample average converges almost surely to the expected value

  $$\bar{\mathbf{X}} \overset{a.s}{\to} \mu \text{ that is } \mathbb{P}\Big(\lim_{n\to\infty} \bar{\mathbf{X}} = \mu\Big) = 1.$$

  This strong law states that the probability of the average of observations converging to the expected values, as the number of trials goes to infinity, is equal to one. The strong law applies to iid random value having an expected value.

There are some technical differences between the strong law and the weak law. The weak law states that for a specified large n, the average $\bar{\mathbf{X}}$ is likely to be near μ. Thus, it leaves open the possibility that $|\bar{\mathbf{X}} - \mu| > \epsilon$ happens an infinite number of times, at very infrequent intervals. The strong law shows that this almost surely will not occur. This doesn't imply the previous statement with probability 1, we have that for any $\epsilon \in \mathbb{R}_+$, the inequality $|\bar{\mathbf{X}} - \mu| < \epsilon$ holds for all large enough n, since the convergence isn't necessarily uniform on the set where it holds.

Chebyshev's inequality: This inequality guarantees that, for a wide class of probability distributions, no more than a certain fraction of value can be more than a certain distance from the mean. Specifically, no more than $\frac{1}{k^2}$ of the distribution's values can be k or more standard deviations away from the mean ie. al least $1 - \frac{1}{k^2}$ of the distribution's values are less than k standard deviations away from the mean.

Let $\mathbf{X}$ be a random variable with finite expected value μ, non-zero finite variance σ^2 then

$$\forall k \in \mathbb{R}_+ \colon\ \mathbb{P}(|\mathbf{X} - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

As an example, using $k = \sqrt{2}$ shows that the probability that the values lie outside the interval $(\mu - \sqrt{2}\sigma, \mu + \sqrt{2}\sigma)$ doesn't exceed $\frac{1}{2}$. For $k = 2$ this value is 25% and when $k = 3$, this value is 12.5%.

In measure-theoretic language, let f be an extended real-valued measurable function defined on X. Then for any real number $t > 0$ and $0 < p < \infty$, then

$$\mu(\{x \in X \ : \ |f(x)| \geq t\}) \leq \frac{1}{t^p} \int_{|f| \geq t} |f|^p \, d\mu.$$

or more generally, if g is an extended real-valued measurable function, non-negative and non-decreasing, with $g(t) \neq 0$

$$\mu(\{x \in X \ : \ f(x) \geq t\}) \leq \frac{1}{g(t)} \int_X g \circ f \, d\mu.$$

Central Limit Theorem: The CLT establishes that, in many situations, the properly normalized sum of independent random variables tends to a normal distribution even if the original variables themselves are not normally distributed. There are several, distinct, formulations of the CLT, as follows

- **Lindeberg-Lévy CLT**: Suppose $\{\mathbf{X}_n\}_{n=1}^{\infty}$ is a sequence of iid random variables with $\mathbb{E}[\mathbf{X}_i] = \mu$ and $\mathbb{V}\text{ar} = \sigma^2 < \infty$. Then, as $n \to \infty$, the random variables $\sqrt{n}(\bar{\mathbf{X}}_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$:

  $$\sqrt{n}(\bar{\mathbf{X}}_n - \mu) \overset{D}{\to} \mathcal{N}(0, \sigma^2).$$

  ie. the cumulative distribution functions of $\sqrt{n}(\bar{\mathbf{X}}_n - \mu)$ pointwise converge to the cumulative distribution functions of $\mathcal{N}(0, \sigma^2)$:

  $$\forall z \in \mathbb{R} \text{ then } \mathbb{P}[\sqrt{n}(\bar{\mathbf{X}}_n - \mu) \leq z] \underset{n\to\infty}{\to} \Phi(\frac{z}{\sigma}),$$

where $\Phi(\frac{z}{\sigma})$ is the standard normal cdf evaluated at z. Equivalenty, the convergence is uniform in z in the sense that

$$\lim_{n\to\infty}\sup_{z\in\mathbb{R}}\left|\mathbb{P}[\sqrt{n}(\bar{\mathbf{X}}_n-\mu)\leq z]-\Phi(\frac{z}{\sigma})\right|=0.$$

- **Lyapunov's CLT**: in this version of the CLT, the sequence of random variables needs not to be identically distributed.

  Suppose $\{\mathbf{X}_n\}_{n=1}^{\infty}$ is a sequence of independent random variables, each with finite expected value μ_i and variance σ_i^2. Let

  $$s_n^2=\sum_{i=1}^{n}\sigma_i^2.$$

  If for some $\delta\in\mathbb{R}_+$, the Lyapunov conditions

  $$\lim_{n\to\infty}\frac{1}{s_n^{2+\delta}}\sum_{i=1}^{n}\mathbb{E}\Big[|\mathbf{X}_i-\mu_i|^{2+\delta}\Big]=0,$$

  is satisfied, then a sum of $\frac{\mathbf{X}_i-\mu_i}{s_n}$ converges to a standard normal random variable, as n goes to infinity

  $$\frac{1}{s_n}\sum_{i=1}^{n}(\mathbf{X}_i-\mu_i)\xrightarrow{D}\mathcal{N}(0,1).$$

  Slutsky's theorem extends the properties of sum, multiplication and division to distribution-convergent sequences of real-valued random variables. This theorem allows us to substitute $s_n\to\sigma^2$.

- **Multivariate CLT**. The previous theorems can be readily extended to $\mathbb{R}^k$-random vectors $\{\mathbf{X}_n\}_{n=1}^{\infty}$ with mean vector $\boldsymbol{\mu}=\mathbb{E}[\mathbf{X}_i]$ and covariance matrix $\mathbf{\Sigma}$, where these random vectors are iid. The multivariate version states that

  $$\sqrt{n}(\bar{\mathbf{X}}_n-\boldsymbol{\mu})\xrightarrow{D}\mathcal{N}_k(0,\mathbf{\Sigma}),$$

  where the covariance matrix is

  $$(\mathbf{\Sigma})_{ij}=\begin{cases}\mathbb{V}\mathrm{ar}(\mathbf{X}_i) & \text{if } i=j\\ \mathbb{C}\mathrm{ov}(\mathbf{X}_i\mathbf{X}_j) & \text{if } i\neq j,\end{cases}.$$

1.3. **Week 3.**

Confidence intervals and CI for normal variables. In general the procedure to construct confidence intervals is to create a pivot or statistic which doesn't depend on the parameter of interest and then solve the probability that the pivot lies between bounds for the parameter.

For small samples, the best confidence intervals can be created using Gosset's t-distribution . To treat Gosset's t-distribution, first we need to discuss the χ^2-distribution .

Chi-squared distribution:

The χ^2-distribution with k degree of freedoms is the distribution of a sum of the squares of k independent standard normal random variables. If $\{\mathbf{X}_n\}_{n=1}^{\infty}$ are independent standard normal random variables $\mathcal{N}(0,1)$, then

$$\text{let } \mathbf{Q} = \sum_{i=1}^{k} \mathbf{X}_i \sim \chi^2_k.$$

In particular, Cochran's theorem establishes that if $\{\mathbf{X}_n\}_{n=1}^{\infty}$ are iid standard normal variables then

$$\sum_{i=1}^{k} (\mathbf{X}_i - \bar{\mathbf{X}}) \sim \chi^2_{k-1} \text{ where } \bar{\mathbf{X}} = \frac{1}{k} \sum_{\mathbf{X}_i}.$$

Suppose that S^2 is the sample variance from a collection of iid $\mathcal{N}(\mu, \sigma^2)$ data, then

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

the χ^2-distribution with $n-1$ degrees of freedom, which has support over $\mathbb{R}_+$, with $\mathbb{E}[\chi^2_n] = n$ and $\mathbb{V}\text{ar}[\chi^2_n] = 2n$. We can use this distribution to create a confidence interval for the variance. Note that if $\chi^2_{n-1,\,\alpha}$ is the α-th quantile of the χ^2-distribution then

$$\begin{aligned} 1 - \alpha = \mathbb{P}\Big(\chi^2_{n-1,\,\frac{\alpha}{2}} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,\,1-\frac{\alpha}{2}}\Big) \\ = \mathbb{P}\Big(\frac{1}{\chi^2_{n-1,\,\frac{\alpha}{2}}} \geq \frac{\sigma^2}{(n-1)S^2} \geq \frac{1}{\chi^2_{n-1,\,1-\frac{\alpha}{2}}}\Big) \\ = \mathbb{P}\Big(\frac{(n-1)S^2}{\chi^2_{n-1,\,1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,\,\frac{\alpha}{2}}}\Big), \end{aligned}$$

so that

$$CI(\sigma^2) = \Big[\frac{(n-1)S^2}{\chi^2_{n-1,\,1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{n-1,\,\frac{\alpha}{2}}}\Big]$$

is a $100(1-\alpha)\%$ confidence interval. The confidence interval for σ can be obtained by square-rooting up the endpoints of the previous CI. Note that, by definition, $\chi^2_{n-1,\,1-\frac{\alpha}{2}} > \chi^2_{n-1,\,\frac{\alpha}{2}}$, with the second one being on the left one's right. This interval relies heavily on the assumed normality. It turns out

$$(n-1)S^2 \sim \text{Gamma}\Big(\frac{(n-1)}{2}, 2\sigma^2\Big),$$

therefore, this can be used to plot a likelihood function for σ^2.

Consider the following example:

Example of χ^2-distribution CI

A recent study on 513 organo-lead manufacturing workers reported an average total brain volume of $1150.315cm^3$ with a standard deviation of $105.977cm^3$. Assuming normality of the underlying measurements, calculate a confidence interval for the popuulation variation in total brain volume.

This can be calculated with the following R-routine

```r
n <- 513
mean <- 1150.315
s2 <- 105.977^2

alpha <- .05
lower_qtile <- qchisq(alpha/2, n-1)
upper_qtile <- qchisq(1-alpha/2, n-1)

```

```
upper_bound <- (n-1)*s2/lower_qtile
lower_bound <- (n-1)*s2/upper_qtile

print(sqrt(lower_bound))
print(sqrt(upper_bound))

In [1]:  99.86484 112.89216
```

or more succinctly

```
n <- 513
mean <- 1150.315
s2 <- 105.977^2

alpha <- .05
qtiles <- qchisq(c(alpha/2, 1-alpha/2), n-1)
ci <- rev((n-1)*s2/qtiles)
sqrt(ci)

In [1]: 99.86484 112.89216
```

Student's t-distribution and CI for normal means. :

t-distribution arises when estimating the mean of a normally distributed population in situations where the sample size is small and the population's standard deviation is unknown (and has to be inferred through a suitable estimator). This distribution plays a central role in a wide number of statistical analyses, including Student's t-test for assessing the statistical significance of the difference between two sample means, the construction of confidence intervals for the difference between two population means. If we take n observations from a normal distribution, then the t-distribution with $\nu = n-1$ degrees of freedom can be defined as the distribution of the location of the sample mean relative to the true mean, divided by the sample standard deviation after multiplying by the standardizing term $\sqrt{n}$. In this way, the t-distribution can be used to construct a confidence interval for the true mean. Said distribution is symmetric and bell shape, but with heavier tails.

Let $\{\mathbf{X}_n\}_{n=1}^{\infty}$ be iid distributions, drawn from the $\mathcal{N}(\mu, \sigma^2)$ distribution. Let

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

be the sample mean and the Bessel-corrected sample variance (an unbiased estimate for the sample variance), respectively. Then, the random variables

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(\mu, \sigma^2), \qquad \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}.$$

This is so since the quantity $\mathbf{V} = (n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2_{n-1}$ by Cochran's theorem. It is readily shown that

$$\mathbf{Z} = (\bar{\mathbf{X}}_n - \mu)\frac{\sqrt{n}}{\sigma} \sim \mathcal{N}(0,1),$$

since the sample mean $\bar{\mathbf{X}}_n$ is normally distributed with mean μ and variance σ^2/n. It also turns out that both $\mathbf{V}$ and $\mathbf{Z}$ are independent. Consequently, the pivotal quantity

$$\mathbf{T} = \frac{\mathbf{Z}}{\sqrt{\frac{\mathbf{V}}{n-1}}} = (\bar{\mathbf{X}}_n - \mu)\frac{\sqrt{n}}{S_n} \sim t_{n-1}$$

which differs from $\mathbf{Z}$ in that the exact standard deviation σ is replaced by the random variable S_n. Notice that the unknown population variance σ^2 doesn't appear in $\mathbf{T}$, which only depends on $\nu = n-1$ but not on μ. In other words

$$t_{df} = \frac{\mathbf{Z}}{\sqrt{\frac{\chi^2}{df}}},$$

which naturally converges to a normal random variable as $df \to \infty$.

Notice that the previous $\mathbf{T}$-random variable can be used as a pivot, therefore it's useful for creating a confidence interval for μ. Thus, let $t_{df,\,\alpha}$ be the α-th quantile of the t-distribution with $df = n-1$ degrees of freedom

$$\begin{aligned} 1-\alpha &= \mathbb{P}\Big(-t_{n-1,\,1-\frac{\alpha}{2}} \leq (\bar{\mathbf{X}}_n - \mu)\frac{\sqrt{n}}{S_n} \leq t_{n-1,\,1-\frac{\alpha}{2}}\Big) \\ &= \mathbb{P}\Big(\bar{\mathbf{X}}_n - t_{n-1,\,1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{\mathbf{X}}_n + t_{n-1,\,1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\Big) \\ &\Rightarrow \mathbb{CI}(\mu) = \bar{\mathbf{X}}_n \pm t_{n-1,\,1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}, \end{aligned}$$

This t-test assumes iid normal data (but it's robust to this assumption) and works well for symmetrically-distributed and mound shaped but for skewed distributions, the t-test performs poorly (since, in these cases, it doesn't make much sense to center the interval at the mean).. Note that, for large degrees of freedom, t-quantiles converge to standard normal quantiles, therefore this random interval converges to the same interval yielded by the CLT.

In R, typing `data(sleep)` brings up the sleep data originally analyzed in Gosset's Biometrika paper, which shows the increase in hours for 10 patients on two soporific drugs. In this case, R treats the data as two independent data groups and not paired. This can be computed and analyzed with the following R-code snippet

```r
data(sleep)

g1 <- sleep$extra[1:10]
g2 <- sleep$extra[11:20]
l1 <- length(g1); l2 <- length(g2)

if (l1 == l2){
    diff <- g2 - g1
} else{
    print("Incompatible list dimensions")
}

mean_diff <- mean(diff)
sample_var <- sd(diff)

if (length(g1) == length(g2)){
    n <- length(g1)
} else if (length(g1) > length(g2)){
    n <- length(g1)
} else if (length(g2) > length(g1)){
    n <- length(g2)
}

CI <- mean_diff + c(-1,1) * qt(.975, n-1) * sample_var / sqrt(n)
```

```
print(CI)

t.test(diff)

In [1]: 0.7001142 2.4598858

In [2]: One Sample t-test

In [3]: data:  diff
In [4]: t = 4.0621, df = 9, p-value = 0.002833
In [5]: alternative hypothesis: true mean is not equal to 0
In [6]: 95 percent confidence interval:
In [7]:  0.7001142 2.4598858
In [7]: sample estimates:
In [8]: mean of x
     1.58
```

The non-central t-distribution : If $\mathbf{X} \sim \mathcal{N}(\mu, \sigma^2)$ and $\mathbf{Y} \sim \chi^2_{df}$, then $\frac{\mathbf{X}/\sigma}{\sqrt{\frac{\chi^2}{df}}}$ is a non-central t-distribution random variable with non-centrality parameter μ/σ. Also note that

$$\text{if } \bar{\mathbf{X}} \sim \mathcal{N}(\mu, \sigma^2) \text{ and } \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$
$$\Rightarrow \frac{\sqrt{n}\bar{\mathbf{X}}}{S} \sim t_{n-1}\Big(\frac{\sqrt{n}\mu}{\sigma}\Big),$$

where μ/σ is the non-centrality parameter. The previous result can then be used to create a likelihood for $\frac{\mu}{\sigma}$, the effect size, with the following R routine (using the sleep trial data), which outputs figure 1

```
data(sleep)

g1 <- sleep$extra[1:10]
g2 <- sleep$extra[11:20]
l1 <- length(g1); l2 <- length(g2)

if (l1 == l2){
    diff <- g2 - g1
} else{
    print("Incompatible list dimensions")
}

mean_diff <- mean(diff)
sample_var <- sd(diff)

if (length(g1) == length(g2)){
    n <- length(g1)
} else if (length(g1) > length(g2)){
    n <- length(g1)
} else if (length(g2) > length(g1)){
    n <- length(g2)
}

non_central_tstat <- sqrt(n) * mean_diff /sample_var
x_vals <- seq(0, 3, length = 1000)
likVals <- dt(non_central_tstat, n-1, ncp = sqrt(n) * x_vals) ## ncp = non-centrality parameter
likVals <- likVals/max(likVals)
plot(x_vals, likVals, type = 'l')

```

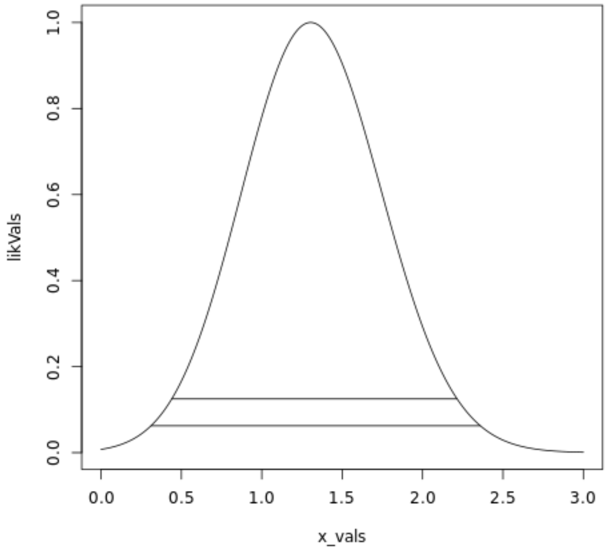


FIGURE 1. Likelihood plot for the effect size $\frac{\mu}{\sigma}$ for the non-central t-distribution .

```
lines(range(x_vals[likVals > 1/8]), c(1/8,1/8))
lines(range(x_vals[likVals > 1/16]), c(1/16,1/16))
```

Profile Likelihoods:

To obtain a likelihood for μ alone, the preferred method is called profiling. The profile for parameter μ_0 is obtained by maximizing the joint likelihood for σ with μ fixed at μ_0, this process being repeated for many values of μ_0.

Consider a joint likelihood with μ fixed at μ_0 given by

$$\propto \prod_{i=1}^{n} \left[(\sigma^2)^{-\frac{1}{2}} \exp\left(- \frac{(x_i - \mu_0)^2}{2\sigma^2} \right) \right].$$

With μ_0 fixed, the ML estimator for σ^2 is $\hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2$[2]. Plugging this into the likelihood returns

$$\left(\sum_{i=1}^{n} \frac{(x_i - \mu_0)^2}{n} \right)^{-\frac{n}{2}} e^{-\frac{n}{2}}.$$

Therefore, up-to multiplicative constants, the profile likelihood is

$$\left(\sum_{i=1}^{n} \frac{(x_i - \mu_0)^2}{n} \right)^{-\frac{n}{2}},$$

which is clearly maximized at $\mu = \bar{\mathbf{X}}$, the same as the ML estimate for μ for the complete likelihood. This profile likelihood is presented at 2 using the sleep dataset with the following R code:

```
data(sleep)

g1 <- sleep$extra[1:10]
g2 <- sleep$extra[11:20]
l1 <- length(g1); l2 <- length(g2)

```

[2] This result is similar in nature to the sample variance, but not exactly since μ_0 is not the sample mean but rather some arbitrary, fixed, number.

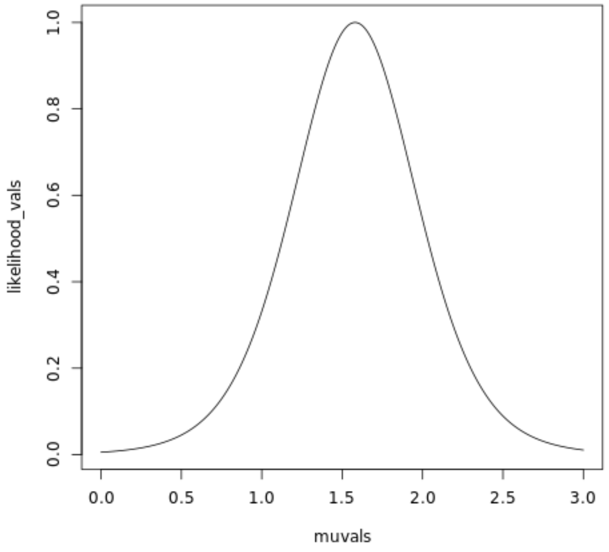


Figure 2. Plot of the profile likelihood for μ of joint likelihood distribution for n iid gaussian $\mathcal{N}(\mu, \sigma^2)$.

```
if (l1 == l2){
    diff <- g2 - g1
    n <- length(g1)
} else{
    print("Incompatible list dimensions")
}

mean_diff <- mean(diff)

muvals <- seq(0, 3, length = 10^3)

prof_lik <- function(mu){
            (sum((diff-mu)^2)/sum((diff-mean_diff)^2))^(-n/2)
            }

likelihood_vals <- sapply(muvals, prof_lik)
plot(muvals, likelihood_vals,type='l')
```

t-confidence intervals. As motivation for developing confidence intervals using the t-distribution , suppose that we desire to compare mean blood pressure between two group in a randomized trial, those who received the treatment versus those who received a placebo. We can't use a paired t-test (basically, taking differences of the data) since the groups are independent and may have different sample sizes. There are, however, different methods for comparing independent groups.

We're interested in constructing a t-interval for the variance, supposing the treated and control group have the same variance,

- Let $\mathbf{X}_1, \cdots, \mathbf{X}_{n_x}$ be iid $\mathcal{N}(\mu_x, \sigma^2)$, the first group's data.
- Let $\mathbf{Y}_1, \cdots, \mathbf{Y}_{n_y}$ be iid $\mathcal{N}(\mu_y, \sigma^2)$, the second group's data.
- Let $\bar{\mathbf{X}}$, $\bar{\mathbf{Y}}$, $\mathbf{S}_x$ and $\mathbf{S}_y$ be the means and standard deviations respectively.

Using the fact that linear combinations of normal random variables are again normal random variables, we know that $\bar{\mathbf{Y}} - \bar{\mathbf{X}} \sim \mathcal{N}(\mu_Y - \mu_X, \sigma^2\Big(\frac{1}{n_x} + \frac{1}{n_y}\Big))$, then the pooled variance estimator

$$\begin{aligned}\mathbf{S}_p^2 &= \frac{(n_x - 1)\mathbf{S}_x^2 + (n_y - 1)\mathbf{S}_y^2}{n_x + n_y - 2} \\ &= \pi\mathbf{S}_x^2 + (1 - \pi)\mathbf{S}_y^2 \text{ where } \pi = \frac{n_x - 1}{n_x + n_y - 2}\end{aligned}$$

is a good estimator for σ^2. It's a mixture of the group variances, placing greater weight on whichever has a larger sample size. Should the sample sizes be equal, the pooled variance estimate is the average of the group variances. As an estimator, it's unbiased since

$$\begin{aligned}\mathbb{E}(\mathbf{S}_p^2) &= \frac{(n_x - 1)\mathbb{E}(\mathbf{S}_x^2) + (n_y - 1)\mathbb{E}(\mathbf{S}_y^2)}{n_x + n_y - 2} \\ &= \frac{(n_x - 1)\sigma^2 + (n_y - 1)\sigma^2}{n_x + n_y - 2}.\end{aligned}$$

Note that the pooled variance estimate is independent of $\bar{\mathbf{Y}} - \bar{\mathbf{X}}$ since $\mathbf{S}_x$ is independent of $\bar{\mathbf{X}}$ and $\mathbf{S}_y$ is independent of $\bar{\mathbf{Y}}$. The sum of two independent χ^2-distribution random variables is again another random variable with its degrees of freedom being the sum of the degrees of freedom of the summands. Therefore, remembering that $\mathbf{V} = (n-1)\frac{\mathbf{S}_n^2}{\sigma^2} \sim \chi_{n-1}^2$ then

$$\begin{aligned}(n_x + n_y - 2)\frac{\mathbf{S}_p^2}{\sigma^2} = (n_x - 1)\frac{\mathbf{S}_x^2}{\sigma^2} + (n_y - 1)\frac{\mathbf{S}_y^2}{\sigma^2} \\ &\sim \chi_{n_x-1}^2 + \chi_{n_y-1}^2 \\ &\sim \chi_{n_x+n_y-2}^2\end{aligned}$$

This is useful to create a pivot, since a t-confidence intervals are constructed by getting a standard normal out of the data and dividing it by its degrees of freedom and the squared root of a χ^2-distribution , in other words

$$\frac{\mathbf{Z}}{\sqrt{\frac{\mathbf{V}}{n-1}}} = (\bar{\mathbf{X}}_n - \mu)\frac{\sqrt{n}}{S_n} \sim t_{n-1}.$$

Then the statistic

$$\frac{\frac{\bar{\mathbf{Y}}-\bar{\mathbf{X}}-(\mu_Y-\mu_X)}{\sigma\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}}{\sqrt{\frac{(n_X+n_Y-2)\mathbf{S}_p^2}{n_X+n_Y-2\sigma^2}}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi_{n_x+n_y-2}^2}{n_x+n_y-2}}},$$

which can be rewritten as

$$\frac{\bar{\mathbf{Y}} - \bar{\mathbf{X}} - (\mu_Y - \mu_X)}{\mathbf{S}_p\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)^{1/2}} \sim t_{n_x+n_y-2}$$

a standard normal (since $\bar{Y} - \bar{X} \sim \mathcal{N}\left(\mu_Y - \mu_X, \sigma^2\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)\right)$) divided by the square root of an independent Chi-squared divided by its degrees of freedom. Thus, this statistic follows a Gosset's t distribution with $n_X + n_Y - 2$ degrees of freedom. This is a special case of an ANOVA test.

Therefore a $(1-\alpha) \times 100\%$-confidence interval for $\mu_Y - \mu_X$ can be written as follows

$$\bar{Y} - \bar{X} \pm t_{n_X+n_Y-2,1-\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}.$$

Notice that the previous random variable follows a non-central t-distribution with non-centrality parameter given by $\frac{(\mu_Y-\mu_X)}{\mathbf{S}_p\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)^{1/2}}$. Then we can use this statistic to create a likelihood for $\frac{\mu_Y-\mu_X}{\sigma}$, a standardized measure of the change in group means.

If we're unwilling to assume equal variances accross the two data groups, we can assert that

$$\bar{\mathbf{Y}} - \bar{\mathbf{X}} \sim \mathcal{N}\Big(\mu_Y - \mu_X, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\Big),$$

since we can't factor out σ^2 out of the square root, then the statistic

$$\bar{Y} - \bar{X} \pm t_{n_X+n_Y-2,1-\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}},$$

approximately follows a Gosset's t-distribution with its degrees of freedom equal to

$$\frac{(\frac{\mathbf{S}_x^2}{n_x} + \frac{\mathbf{S}_y^2}{n_y})^2}{\left(\frac{\mathbf{S}_x^2}{n_x}\right)^2 \frac{1}{n_x-1} + \left(\frac{\mathbf{S}_y^2}{n_x}\right)^2 \frac{1}{n_y-1}}.$$

The confidence interval can be readily computed using the appropriate t-quantile.

Jackknife. The jackknife is a tool for estimating standard errors and the bias of estimators. Both the jackknife and the bootstrap involve resampling data, ie. repeatedly creating new data sets from the original data.

The jacknife deletes each observation and calculates an estimated based on the remaining $n-1$ of them. It uses this collection of estimates to do things like estimating bias and standard errors. Note that estimating the bias and having a standard error are not relevant quantities to compute a sample mean, which we know are unbiased estimates of population means and what their standard error are.

Consider the jackknife for univariate data. Let $\{\mathbf{X}_n\}_{n=1}^{\infty}$ be a collection of data used to estimate a parameter θ, with $\hat{\theta}$ being the estimate based on the full data-set. Let $\hat{\theta}_i$ be the estimate of θ obtained by deleting the i-th observation and finally let $\bar{\theta} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_i$. Then the jackknife estimate for the bias is

$$(n-1)(\bar{\theta} - \hat{\theta}),$$

ie. how far the average delete-one estimate is from the actual estimate, with the jackknife standard error is

$$\Big[\frac{n-1}{n} \sum_{i=1}^{n} (\bar{\theta} - \hat{\theta})^{\frac{1}{2}}\Big],$$

ie. the deviance of the delete-one estimates from the average delete-one estimate.

Example: Using the Jackknife

Consider the data-set of 630 measurements of gray matter volume for workers from a lead manufacturing plant. The median gray matter volume if around 589 cubic centimetres and we wish to estimate the bias and standard error of the median.

The general procedure can be implemented as an algorithm in R with the following routine:

```r
n <- length(gmVol)
theta <- median(gmVol)

jk <- sapply(1:n,
             function(i) median(gmVol[-i])
             )
thetaBar <- mean(jk)
biasEst <- (n-1) * (thetaBar - theta)
seEst <- sqrt((n-1) * mean((thetaBar - theta)^2))
```

or using the boostrap package

```r
library(bootstrap)
out <- jackknife(gmVol, median)
out$jack.se
out$jack.bias
```

Both methods yield an estimated bias of 0 and a standard error of 9.94. In general, the jackknife estimate of the bias for the median is always 0 when the number of observations is even.

It has been proven that the jackknife is a linear approximation to the bootstrap. Generally it's not recommended to use the jackknife for sample quantiles like the median, since it has poor properties.

Another interesting way to think about the jackknife uses pseudo observations, defined as

$$\text{Pseudo Obs } = n\hat{\theta} - (n-1)\hat{\theta}_i,$$

in other words these are "whatever observation i contributes to the estimate of θ". Note that when θ is the sample mean, the pseudo observations are the data themselves.

Then the sample standard error of these observation is the previous jackknife estimated standard error. The mean of these observations is a bias-corrected estimate for θ.

Bootstrapping. The bootstrap is a tremendously useful tool for constructing confidence intervals and calculating standard errors for difficult statistic.

Suppose that a certain statistic, for example the median, estimates some population parameter, but it's sampling distribution is not known. The bootstrap principle suggests using the distribution defined by the data to approximate its sampling distribution. Said data-induced distribution is discrete and puts probability $\frac{1}{N}$ on every one of the N data points with its mean being the sample mean.

In practice, the bootstrap principle is always carried out by simulations. The general procedure follows by first simulating complete data sets from the observed data with replacement[3] (this is approximately drawing from the sampling distribution of that statistic, at least as far as the date is able to approximate the true population distribution), then calculating the statistic for each simulated data set. Then, we use the simulated statistics to either define a confidence interval or take the standard deviation to calculate a standard error.

[3] Note that sampling without replacement would wind up to be a permutation of the original data set.

Sampling with replacing is exactly drawing iid samples from the empirical distribution, which places probability $\frac{1}{N}$ on each data point. Consider the previous' section example:

Example: Using the Jackknife

Consider the data-set of 630 measurements of gray matter volume for workers from a lead manufacturing plant. The median gray matter volume if around 589 cubic centimetres and we wish to estimate the bias and standard error of the median.

The bootstrap procedure for calculating the median from a data set of n observations is

- Sampling n observations with replacement from the observed data resulting in one simulated data set,
- taking the median of the simulated data set,
- repeating these two steps B times, resulting in B medians.
- These medians are approximate draws from the sampling distribution of the median of n observations, they are exact draws from the sampling distribution of the median of N observations from the distribution of the observed data, therefore we can
  - draw a histogram of them,
  - calculate their standard deviations to estimate the standard error of the median,
  - or take the 2.5-th and 97.5-th percentiles as a confidence interval for the median. This is a bootstrap confidence interval.

Numerically, this can be done in R:

```r
B <- 1000
n <- length(gmVol)
resamples <- matrix(sample(gmVol,
                           n*B,
                           replace=TRUE),
                    B,n)
medians <- apply(resamples, 1, median)
sd(medians)
```

or by using the bootstrap library

```r
library(boot)
stat <- function(x, i) {median(x[i])}
boot.out <- boot(data = gmVol,
                 statistic = stat,
                 R = 1000)
boot.ci(boot.out)
```

The bootstrap is non-parametric and work best for large samples. Further details can be found in ”An introduction to the Bootrstrap” by Efron and Tibshirani.

1.4. **Week 4.**

Binomial proportions. :

Let $\mathbf{X} \sim$ Binom(n, p) we know that

- $\hat{p} = \frac{\mathbf{X}}{n}$ is the maximum likelihood estimator for p, the sample proportion of successes.
- $\mathbb{E}[\hat{p}] = p$,
- $\mathbb{V}\text{ar}[\hat{p}] = \frac{p(1-p)}{n}$,

- Since $\hat{p}$ is an average of Bernoulli trials, the Central Limit Theorem holds, therefore $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$ follows a normal distribution for large n.

The latter fact leads to the Wald interval for p

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

which performs poorly. Asymptotically, coverage at α-th level is guaranteed, however in practice coverage probability varies wildly, sometimes being quite low for certain values of n even when p is not near the boundaries. For example, when $p = .5$ and $n = 40$, the actual coverage of a 95% interval is only 92%. When p is small or large, coverage can be quite poor even for extremely large values of n. For example, when $p = .005$ and $n = 1876$ the actual coverage rate of a 95% interval is only 90%.

A simple fix for the problem is to add two successes and two failures, which is yet again a random variable, and to treat the latter as if it were the data. That is, let $\tilde{\mathbf{p}} = \frac{\mathbf{X}+2}{n+4}$. The Agresti-Coull interval is

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

When p is either very large or very small, the distribution of $\hat{p}$ is skewed and it doesn't make sense to center the interval at the MLE, but rather at another artificially created center with pseudo-observations, pulling the center of the interval towards .5. This interval is the inversion of the hypothesis testing technique.

Binomial proportions example:

Suppose that in a random sample of an at-risk population of 13 out of 20 subjects had hypertension and we'd like to estimate the prevalence of hypertension in this population.

- $\hat{p} = .65$, $n = 20$,
- $\tilde{p} = .65$, $\tilde{n} = n + 4 = 20$,
- $z_{975} = 1.96$,
- Wald interval [.44, .86],
- Agresti-Coull interval [.44, .82],
- $\frac{1}{8}$ likelihood interval [.44, .84].

Bayesian Analysis. :

Bayesian statistics posits a **prior** on the parameter of interest, a probability distribution that represent our beliefs on that parameter. All inferences are then performed on the distribution of the parameter given the data, the so called **posterior**. In general

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}.$$

Therefore the likelihood is the factor by which our prior beliefs are updated to produce conclusions in light of the data.

Our binomial data is discrete, only taking values between 0 and 1, but the proportion to be estimated is continuous, $p \in \Theta \sim \mathbb{R}_{[0,1]}$. Thus we need a continuous distribution, with a lower bound at 0 and an upper bound at 1. The beta distribution is the default prior for parameters between 0 and 1. The beta density depends on two parameters, α and β, and its pdf is given by

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1} \text{ for } 0 \le p \le 1,$$

where $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ arises as a normalization constant since $\int_{p\in\Theta} dpp^{\alpha-1}(1-p)^{\beta-1} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Its mean is $\frac{\alpha}{\alpha+\beta}$ and its variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. The uniform density is the special case where $\alpha = \beta = 1$.

Suppose that we chose values of α and β so that the beta prior is indicative of our degrees of belief regarding p in the absence of data. Then, according to Bayes' rule

$$\text{Posterior } \propto \text{ Likelihood } \times \text{ Prior.}$$

Therefore, up-to to p-independent multiplicative factors

$$\begin{aligned}\text{Posterior } &\propto p^x(1-p)^{n-x} \times p^\alpha(1-p)^{\beta-x} \\ &\propto p^{x+\alpha-1}(1-p)^{n-x+\beta-1},\end{aligned}$$

which is another beta density with parameters $\tilde{\alpha} = x + \alpha$ and $\tilde{\beta} = n - x + \beta$. Then, we can calculate the posterior mean given the data is

$$\begin{aligned}\mathbb{E}[p|\mathbf{X}] &= \frac{\tilde{\alpha}}{\tilde{\alpha}+\tilde{\beta}} \\ &= \frac{x+\alpha}{x+\alpha+n-x+\beta} \\ &= \frac{x+\alpha}{n+\alpha+\beta} \\ &= \frac{x}{n}\cdot\frac{n}{n+\alpha+\beta} + \frac{\alpha}{\alpha+\beta}\cdot\frac{\alpha+\beta}{n+\alpha+\beta} \\ &= \text{MLE } \times \pi + \text{ Prior Mean } \times (1-\pi),\end{aligned}$$

where $\pi = \frac{n}{n+\alpha+\beta}$ and where we used the fact that the posterior is the distribution of the parameter given the data, the likelihood is the probability of the data given the parameter and the prior is the probability of the parameter disregarding the data. So, the previous equation is an average of the MLE and the prior mean, but each without .5 probability each. Note that when $n >> 1$, then $\pi >> 1$ and the MLE term dominates. Therefore, as the sample size gets bigger, the data means more. On the contrary, as $\alpha, \beta >> 1$ and n remains constant, the prior mean term dominates.

The posterior variance is

$$\begin{aligned}\mathbb{V}\text{ar}[p|\mathbf{X}] &= \frac{\tilde{\alpha}\tilde{\beta}}{(\tilde{\alpha}+\tilde{\beta})^2(\tilde{\alpha}+\tilde{\beta}+1)} \\ &= \frac{(x+\alpha)(n-x+\beta)}{(n+\alpha+\beta)^2(n+\alpha+\beta+1)}.\end{aligned}$$

Let $\tilde{p} = \frac{x+\alpha}{n+\alpha+\beta}$ and let $\tilde{n} = n + \alpha + \beta$ then

$$\mathbb{V}\text{ar}[p|\mathbf{X}] = \frac{\tilde{p}(1-\tilde{p})}{\tilde{n}+1},$$

which is very similar (but not quite) to the variance of a binomial random variable.

Consider now the previous example where $x = 13$ and $n = 20$, with a uniform prior of $\alpha = \beta = 1$. Then, the posterior is proportional to

$$p^{x+\alpha-1}(1-p)^{n-x+\beta-1} = p^x(1-p)^{n-x},$$

that is, for the uniform prior, the posterior itself is the likelihood. Consider the instance where $\alpha = \beta = 2$, the posterior is

$$p^{x+\alpha-1}(1-p)^{n-x+\beta-1} = p^x(1-p)^{n-x+1}$$

The particular case of $\alpha = \beta = .5$ is called Jeffrey's prior.

The posterior, in Bayesian analysis, is the distribution of the parameter given the data and is the object of study of Bayesian analysis. One way to summarize data was to use confidence intervals in frequentist statistics. However, said confidence intervals lies on the idea of obtaining similar results in fictitious repetitions of experiments, idea not compatible with Bayesian statistics. Therefore, we define a Bayesian credible interval is the Bayesian analog of the confidence interval. A 95% credible interval $[a, b]$ would satisfy

$$\mathbb{P}(p \in [a, b] | x) = .95,$$

ie. it's the random interval which satisfies that the probability of the parameter lying in the interval, given the data, is 95%. Higher values of the posterior represent better supported values of the parameter, so the best credible intervals chop off the posterior with a horizontal line, similarly to likelihoods. These are called highest posterior density intervals. This can be computed in R by using the `binom.bayes(13,20,type="highest")` within the `binom` library, for a default credible level of 95% and with the default prior being Jeffrey's prior.

It's important to understand the correct interpretation of both confidence intervals and Bayesian credible intervals.

- Formally, a confidence interval is constructed in such a way that in repeated (fictitious) independent experiments, 95% of the intervals obtained would contain p.

- A (1/8-th) likelihood interval represents plausible values for p in the sense that for each point in this interval, there is no other point that is more than 8 times better supported given the data.

- The Jeffrey's Bayesian credible interval's interpretation is straightforward, it's the random interval constructed in such a way that the probability that p lies in said interval is 95%, where probability must be interpreted according to the Bayesian approach.

Logs and the Geometric Mean: Taking logs of data is a well-known and widely used technique in data analysis, useful to correct for right skewness, when considering rations, in setting where errors are feasibly multiplicative, such as when dealing with concentration or rates, when analysing data with high orders of magnitude and counts/frecuencies are also logged.

The sample geometric mean of a data set $\{\mathbf{X}_n\}_{n=1}^{\infty}$ is

$$\Big(\prod_{i=1}^{n} \mathbf{X}_i\Big)^{\frac{1}{n}},$$

provided the $\mathbf{X}_i$ are positive, the log of the geometric mean is then

$$\frac{1}{n}\sum_{i=1} n \log \mathbf{X}_i.$$

Then, the log of the geometric mean is an average, where the law of large numbers and the central limit theorem apply. The log of geometric mean is the arithmetic mean of the log of observations. Note that the geometric mean is always less than or equal to the sample arithmetic mean. The geometric mean is often used when the $\mathbf{X}_i$ are all multiplicative. Consider the following example

Geometric mean Example

Suppose that in a population of interest, the prevalence of a disease rose 2% one year, then fell 1% the next, then rose 2%, then rose 1%; since these factors act multiplicatively it makes sense to consider the geometric mean:

$$(1.02 \times .99 \times 1.02 \times 1.01)^{\frac{1}{4}} = 1.01,$$

for a 1% geometric mean increase in disease prevalence.
Note that 1.01 is the constant factor by which you would need to multiply the initial prevalence each year to achieve the same overall increase in prevalence over a four year period. In contrast, the arithmetic mean is the constant factor by which you would need to add each year to achieve the same total increase (1.02+.99+1.02+1.01). In this case, the product and hence the geometric mean make more sense than the arithmetic mean.

Another way to interpret the differences between the geometric and arithmetic means is to consider the following example. Let a and b be the lengths of the sides of a rectangle, then

- The arithmetic mean $\frac{a+b}{2}$ is the length of the sides of the square that has the same perimeter,
- and the geometric mean $\sqrt{ab}$ is the length of the sides of the square that has the same area.

Then if we're interested in perimeters (adding) use the arithmetic mean; if we're interested in areas (multiplying), we use the geometric mean.

Now, according to the law of large numbers, the log of the geometric mean converges to $\tilde{\mu} = \mathbb{E}[\log(\mathbf{X})]$. Therefore, the geometric mean converges to $\exp \mathbb{E}[\log(\mathbf{X})] = e^{\tilde{\mu}}$, the population geometric mean, which is not the population mean on the natural scale,

$$\exp \mathbb{E}[\log(\mathbf{X})] \neq \mathbb{E}[\exp \log(\mathbf{X})] = \mathbb{E}[\mathbf{X}].$$

Note that if the distribution of $\log \mathbf{X}$ is symmetric, then

$$.5 = \mathbb{P}(\log \mathbf{X} \leq \tilde{\mu}) = \mathbb{P}(\mathbf{X} \leq e^{\tilde{\mu}}),$$

since $\log \cdot$, $\exp(\cdot)$ and $\mathbb{E}[\cdot]$are monotonically increasing functions. This is one of the reasons for taking logs of data. Therefore, for log-symmetric distributions, the geometric mean is estimating the median.

If we use the central limit theorem to create a confidence interval for the log-measurements, the interval is estimating $\tilde{\mu}$, the expected value of the log-measurements. By exponentiation of the endpoints of the interval, the interval estimates $e^{\tilde{\mu}}$, the population geometric mean. Should the logged data be symmetric, then $e^{\tilde{\mu}}$ is the population median. This is a useful technique for paired data, when their ratio, rather than their difference, is of interest.

Consider now the following example

Example

Consider a paired design experiment comparing systolic blood pressure (SBP) for people taking oral contraceptives for users and matched controls. We're interested in analysing the logs of the ratios of test subject group and control group since the ratios give us information about increases and decreases of the SBP.

- The geometric mean ratio is 1.04 (ie. a 4% increase in SBP for the OC users),
- The T-interval on the difference of the log scale measurements is $\mathbb{C}(\log(\text{ratios})) = [0.010, 0.067]\log(\text{mmHg})$.
- Exponentiating yields the confidence interval $\mathbb{C}((\text{ratios})) = [1.010, 1.069]\text{mmHg}$.

Hence, we can confirm an increase in SBP in the test subjects group compared to the control group.

Consider now an experiment involving testing two independent groups, logging the individual data points and creating a confidence interval for the differences in the log means, if the data is symmetric in the log-scale, then it's also equal to a ratio in the population medians. Then exponentiating the endpoints of this interval yields the interval for the ratio of the population geometric means, $\frac{e^{\tilde{\mu}_1}}{e^{\tilde{\mu}_2}}$.

A random variable is log-normally distributed if its log is a normally distributed random variable. Note that log-normal random variables are not log of normal random variables. Formally, $\mathbf{X} \sim \mathcal{LN}(\mu, \sigma^2)$ if $\log \mathbf{X} \sim \mathcal{N}(\mu, \sigma^2)$. Note that if $\mathbf{X} \sim \mathcal{N}(\mu, \sigma^2)$ then $\mathbf{Y} = e^{\mathbf{X}}$ is log-normal. The log-normal density is given by

$$\frac{1}{\sqrt{2\pi}} \frac{\exp[-(\log x - \mu)^2/(2\sigma^2))]}{x},$$

with support over $x \in \mathbb{R}_{(0,\infty)}$, with a mean of $e^{\mu+\frac{\sigma^2}{2}}$, variance of $e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$ and median e^{μ}.

Let $\{\mathbf{X}_n\}_{n=1}^{\infty}$ be sequence of log-normal(μ, σ^2) random variables, then $\{\mathbf{Y}_j\}_{j=1}^{n} | \mathbf{Y}_k = \log \mathbf{X}_k$ are normally distributed with mean μ and variance σ^2. Therefore, we can create a Gosset's t-confidence interval. If μ is the log of the median of the $\mathbf{X}_i$, then e^{μ} gives the median on the original scale and it also population geometric mean. Assuming log-normality, exponentiating t-confidence intervals for the difference in two log means, again estimates ratios of geometric means.

Consider the following example of log-normal random variables.

Example of log-normal random variables

Consider an experiment of gray matter volumes, test subjects being divided amongst two groups, young group and old group. Then we take GM volumes for young and old groups by logging them, the results being

$$\mathbb{CI}(\text{log old}) = [13.24, 13.27] \log \text{cm}^3 \text{ and } \mathbb{CI}(\text{young}) = [13.29, 13.31] \log \text{cm}^3.$$

By exponentiation, this yields

$$\mathbb{CI}(\text{old}) = [564.4, 577.5]cc \text{ and } \mathbb{CI}(\text{young}) = [592.0, 606.9]cc.$$

Performing a two group t-interval on the difference of the logged measurement yields $\mathbb{CI}(\text{log young - log old}) = [0.032, 0.066] \log \text{cm}^3$, which exponentiated yields $\mathbb{CI}(e^{\text{log young}}/e^{\text{log old})} = [1.032, 1.068]cm^3$.

2. Part 2

2.1. **Week 1.**

2.1.1. *Hypothesis testing.* Hypothesis testing is concerned with making decisions using data. We label a null hypothesis as H_0. The null hypothesis is assumed true and statistical evidence is required to rejected in favour of an alternative hypothesis.

For example, a respiratory disturbance index of more than 30 events/hour is considered evidence of severe sleep disorder breathing. Suppose that in a sample of 100 overweight subjects with other risk factors for sleep disordered breathing at a sleep clinic, the mean RDI was 32 events/hour with a standard deviation of 10 events/hour. We might want to test the hypothesis that

- $H_0 : \mu = 30$
- $H_a : \mu > 30$

where μ is the population mean RDI. The alternative hypothesis are typically of the form $<, >$ *or* $\neq$. Note that there four possible outcome of our statistical decision process

Statistical decision process		
Truth	H_0	H_a
H_0	Correctly accept null	Type I error
H_a	Type II error	Correctly reject null

- The type I error rate is a false positive *ie.* the mistaken rejection of a (true) null hypothesis.
- The type II error rate is a false negative *ie.* the mistaken acceptance of a (false) null hypothesis, we failed to reject a false H_0.

Let α denote the type I error rate, the probability of rejecting the null hypothesis when, in fact, then null hypothesis is correct. We'd like to minimise this kind of error. Considering our previous example, a reasonable strategy would reject the null hypothesis if $\bar{X}$ was larger than some constant $C \in \mathbb{R}_+$. Typically, C is chosen so that the probability of a Type I error α is 0.05.

For example, the probability of a Type I error and according to the Central Limit Theorem we have

$$\begin{aligned} .05 &= \mathbb{P}\Big(\bar{X} \geq C \Big| \mu = 30\Big) \\ &= \mathbb{P}\Big(\frac{\bar{X} - 30}{10/\sqrt{100}} \geq \frac{C - 30}{10/\sqrt{100}} \Big| \mu = 30\Big) \\ &= \mathbb{P}\Big(Z \geq \frac{C - 30}{10/\sqrt{100}}\Big) \end{aligned}$$

Hence $\frac{C-30}{1} = 1.645$ implying $C = 31.645$. Now, since our mean is 32 we reject the null hypothesis. We can plot a the 95th percentile of the standard normal distribution with the following R routine:

```r
xval <-  seq(-3.2, 3.2, length=1000)
yval <- dnorm(xval)
plot(xval, yval, type="l", axes=TRUE, frame=FALSE, lwd=3, xlab="", ylab="")
x <- seq(qnorm(.95), 3.2, length=1000)
polygon(c(x,rev(x)), c(dnorm(x), rep(0, length(x))), col="salmon")
text(mean(x), mean(dnorm(x))+0.2, "5%", cex=2)
text(qnorm(.95), .01, "1.645",cex=2),
```

which outputs plot 3.

In general, we don't convert C back to its original scale. We would just reject it because the Z-score, how many standard error units the sample mean is above the hypothesised mean, is greater than 1.645 *ie.*

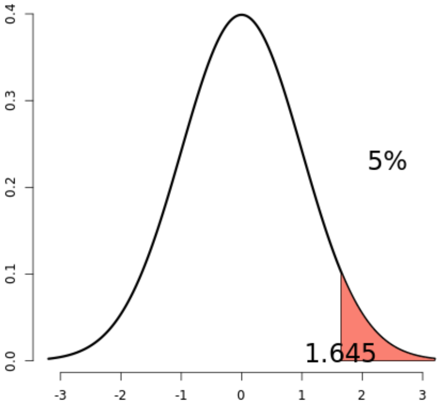


FIGURE 3. R-generated plot of the standard normal distribution $\mathcal{N}(0,1)$, in the [-3.2, 3.2] range. The shaded polygon represents the 95th percentile (which is given by $x = 1.645$) of a standard normal distribution.

$$\frac{32-30}{10/\sqrt{100}} = 2 > 1.645.$$

Our mean is two standard error units away from the hypothesised mean. We can codify these rules for a normal Z test, assuming the data is gaussian or that the Central Limit Theorem is a good enough approximation to apply, as follows:

Rules for a normal Z-test

Let $H_0 : \mu = \mu_0$ be the null hypothesis, then let

- $H_1 : \mu < \mu_0$
- $H_2 : \mu \neq \mu_0$
- $H_3 : \mu > \mu_0$

and let the test statistic be

$$TS = \frac{\bar{X} - \mu_0}{S/\sqrt{a}},$$

ie. expressing the mean in standard error units), where $S/\sqrt{a}$ is the standard error. Thus, we reject the null hypothesis when

- $H_1 : TS \leq -z_{1-\alpha}$ *ie.* we reject H_0 in favour of H_1 if our sample mean is enough below μ_0,
- $H_2 : |TS| \geq z_{1-\alpha/2}$ *ie.* we reject in favour of H_2 if our sample mean is enough too different from μ_0 (either too large or too small)[a],
- $H_3 : TS \geq z_{1-\alpha}$ *ie.* we reject in favour of H_3 if our sample mean is enough above μ_0.

[a] In this case, we look at the $(1-\alpha/2) \times 100\%$ error rate, we divide the probability of a type I error into half of it being accidentally rejected because of sample mean being too large and the other half of it being accidentally rejected because the sample mean is too small.

The region of TS values for which the null hypothesis is rejected is the rejection region. In the case of H_1, the upper normal quantile and above is the rejection region, for H_2 it's the negative quantile and below or the upper quantile and above and finally, in the case of H_3, the normal quantile and below is the rejection region. We can graphically represent the second case, the case of a two-sided tail test, with the following R routine

```
xval <-  seq(-3.2, 3.2, length=1000)
```

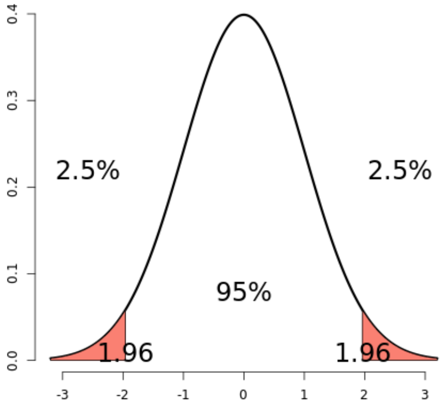


Figure 4. Two sided tail test presenting a standard normal $\mathcal{N}(0, 1)$ distribution. We're going to reject H_0 if our test statistic is above 1.96 -which has a 2.5% chance under the null hypothesis- and we're going to reject H_0 if our test statistic is below -1.96 -which also has a 2.5% chance under the null hypothesis-. Thus the union of those two events has a 5% chance under the null hypothesis.

```r
yval <- dnorm(xval)

plot(xval, yval, type = "l", axes=TRUE, frame=FALSE, lwd = 3, xlab="", ylab= "")
x <- seq(qnorm(.975), 3.2, length = 100)
polygon(c(x, rev(x)), c(dnorm(x), rep(0,length(x))), col="salmon")
text(mean(x), mean(dnorm(x))+0.2, "2.5%", cex=2)
text(qnorm(.975), .01, "1.96",cex=2)

x <- seq(-3.2, qnorm(0.025), length=100)
polygon(c(x, rev(x)), c(dnorm(x), rep(0,length(x))), col="salmon")
text(mean(x), mean(dnorm(x))+0.2, "2.5%", cex=2)
text(qnorm(.025), .01, "1.96",cex=2)
text(0, dnorm(0)/5, "95%",cex=2)
```

In hypothesis testing, we fix the α (type I error value) to be low. So if we reject the null hypothesis, either our model is wrong or there is a low probability that we made and error. We haven't fixed the probability of a type II error β, therefore we tend to say "fail to reject H_0" rather than accepting H_0. In general, less is known about the type II error. Statistical significance is not the same as scientific significance.

2.1.2. *Two sided tests.* The Z-test requires the assumption of the Central Limit theorem and for n to be large enough for the CLT to apply. If n is mall, then a Gossett's T test is performed exactly in the same way with the normal quantiles being replaced by the appropriate Student's T quantiles and $n-1$ degrees of freedom.

The probability of rejecting a false null hypothesis is called **power**. The statistical power of a dichotomic (binary) hypothesis test is the probability that the test correctly rejects the null hypothesis H_0 when a specific alternative hypothesis H_1 is true[4]. A high value of power is good thing, we want to reject a false null hypothesis but it's not an easy to manipulate quantity in an experiment. One way to combat this issue is, prior to conducting the study, to do a power calculation of the sample size to obtain a certain level of power using guesses for the standard deviation and the hypothesised significant effect.

For example, suppose that $n = 16$ (rather than 100) in our previous example. Instead of using a Z-test, we use a Student's T-test. Then

$$.05 = \mathbb{P}\Big(\frac{\bar{X} - 30}{s/\sqrt{16}} \geq t_{1-\alpha,15}\Big|\mu = 30\Big)$$

where $s/\sqrt{16}$ is the estimated standard error, where $t_{1-\alpha,15}$ is t-quantile. Then our test statistic is now $\sqrt{16}\frac{32-30}{10} = .8$, while the critical value is $t_{1-\alpha,15} = 1.75$. Thus, we fail to reject the null hypothesis. Remember that T has a heavier tail distribution than the normal distribution.

We are now interested in a two sided test. We are interested in studying if our hypothesised mean is significantly larger than 30 or significantly smaller than 30. That is, we want to test he alternative hypothesis $H_a : \mu \neq 30$. Then

$$\alpha = \mathbb{P}\Big(\Big|\frac{\bar{X} - 30}{s/\sqrt{16}}\Big| > t_{1-\alpha/2,15}\Big|\mu = 30\Big),$$

that is: we'll reject H_0 is the test statistic (0.8) is either too large or two small with the critical value being calculated using $\alpha/2$. In our example, this critical value is 2.13 so we fail to reject it[5] The previous calculation can be understood with the following R routine

```
xval <-  seq(-4, 4, length=1000)
yval <- dt(xval, 15)
plot(xval, yval, type = "l", axes=TRUE, frame=FALSE, lwd = 3, xlab="", ylab= "")
x <- seq(qt(.975, 15), 4, length=100)
polygon(c(x, rev(x)), c(dt(x,15), rep(0,length(x))), col="salmon")
text(mean(x), mean(dt(xval, 16-1))+0.2, "2.5%", cex=2)
text(qt(.975,15), .01, "2.13",cex=2)
x <- seq(-3.2, qt(.025,16), length=100)
polygon(c(x, rev(x)), c(dt(x,15), rep(0,length(x))), col="salmon")
text(mean(x), mean(dt(xval, 16-1))+0.2, "2.5%", cex=2)
text(qt(.025,15), .01, "2.13",cex=2)
text(0, dt(0,15)/5, "95%", cex=2)
```

which outputs figure 5.

[4] Remember that the probability of rejecting a true null hypothesis is the type I error rate and is set to be small whilst the probability of failing to reject (this is the mistaken acceptance) a false null hypothesis is called a type II error rate. The power is $1 - \beta$.

[5] We note that, if the data are iid Gaussians, then $\frac{X-\mu}{s/\sqrt{n}} \sim t_{n-1}$

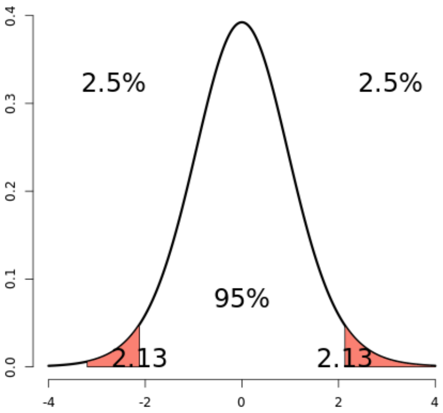


Figure 5. We calculate our normalised test statistic $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ and the probability that the absolute value of said statistic is bigger than 2.13 *ie.* the probability that the statistic is too large positive than 2.13 is 2.5% or too small negative than -2.13 is 2.5%. In our case, the probability that the test statistic lies in the rejection region is 5%. For the T distributionwith 15 degrees of freedom, the 97.5th quantile is 2.13 and the 2.5th quantile is -2.13.

2.1.3. *Confidence intervals and P values.* Consider testing $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$. Take the set of all possible values for which we fail to reject H_0, this set is a $(1-\alpha) \times 10\%$ confidence interval for μ, these are the value for μ that are supportable as null hypotheses (those are reasonable numbers for μ). This rule works both ways, if a $(1-\alpha) \times 10\%$ contains μ_0, then we fail to reject H_0.

Consider that we don't reject H_0 if our test statistic behaves as follows:

$$\left|\frac{\bar{X}-\mu}{s/\sqrt{n}}\right| \leq t_{1-\alpha/2,n-1},$$

which implies

$$|\bar{X}-\mu| \leq t_{1-\alpha/2,n-1} s/\sqrt{n}$$

which, in turn, implies that

$$\bar{X} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu_0 < \bar{X} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}.$$

This states that if μ_0 lies inside the confidence interval, the we've failed to reject H_0 and this argument is reversible. This establishes a duality between confidence intervals and two-sided hypotheses tests. If we create a 95% confidence interval, it conveys more information than the result of a hypothesis test because we can do the hypothesis test and gives a sense of what values for μ_0 are well supported, helping to reduce the gap between statistical significance and scientific significance.

In our previous example, we rejected the one-sided test when $\alpha = .05$, would we reject it if .01 or .001?

The smallest value for α that you still reject the null hypothesis is called the **attained significance level**. This equivalent, but philosophically different, from the **P-value**. The P-value (same number but different concept) is the probability, under the null hypothesis, of obtaining evidence as extreme or more extreme than would be observed by chance alone, where chance is governed by the null distribution. If the P-value is small, then either H_0 is true and we have observed a rare event, given that the null hypothesis is true, or H_0 is false. It quantifies whether or not getting a test statistic as or more extreme than you observed was rare under the null hypothesis. If it's rare, then that casts some doubt on the veracity of the null hypothesis.

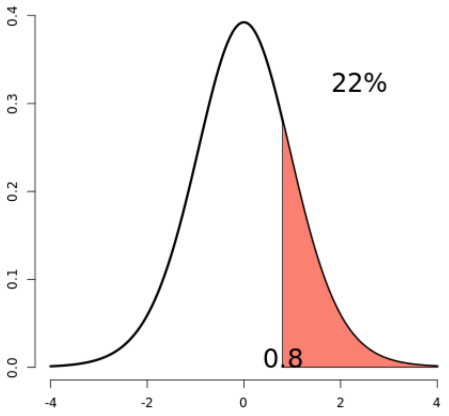


FIGURE 6. Our statistic $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ turned out to be .8. The probability of being .8 or larger for a Student's T distribution with 15 degrees of freedom is 22%. Since this number is larger than $\alpha = 5\%$, we would fail to reject H_0.

In our example, the T statistic was 0.8. What's the probability of getting a T statistic as large as 0.8? This can be computed with the following R code

```
pt(0.8, 15, lower.tail = FALSE)

In [1]:   0.218099
```

where `pt` stands for T probability, `lower.tail = FALSE` indicates we want a value above 0.8, not lower. This works out to be $P = 22\% > \alpha = 5\%$, which is entirely reasonable because we failed to reject the null hypothesis.

- If the P-value is larger than α, we fail to reject H_0.
- If the P-value is smaller than α, we reject H_0

This can be understood via the following R routine

```
pt(0.8, 15, lower.tail=FALSE)
xval <-  seq(-4, 4, length=1000)
yval <- dt(xval, 15)
plot(xval, yval, type = "l", axes=TRUE, frame=FALSE, lwd = 3, xlab="", ylab= "")
x <-  seq(.8, 4, length=100)
polygon(c(x, rev(x)), c(dt(x,15), rep(0,length(x))), col="salmon")
text(mean(x), mean(dt(xval, 16-1))+0.2, "22%", cex=2)
text(.8, .01, "0.8",cex=2)
```

which outputs figure 6.

By reporting a P-value, the reader can perform a hypothesis test at whatever α level he or she chooses. Again, is the P-value is less than α, we reject the null hypothesis.

For two sided hypothesis test, double the smaller of the two one sided hypothesis test P-values.

Criticisms of the P value:

- P values only consider significance, unlike confidence intervals making it harder to distinguish practical significance from statistical significance.
- Absolute measures of the rareness of an event are not good measures of the evidence for or against a hypothesis.

2.1.4. *Power.* Power is the probability of rejecting the null hypothesis when it's false. And, as the name suggests, it's a desirable quality, the more power, the more reliable the test is. A type II error is failing to reject the null hypothesis when it's false. Thus, by their definitions, power + type II error rate (β) sum to one. Thus we define Power $= 1 - \beta$.

Consider the example involving RDI. The null hypothesis is $H_0 : \mu = 30$ versus $H_a : \mu > 30$. Then power is calculated under the alternative hypothesis, as follows

$$\mathbb{P}\Big(\frac{\bar{X} - 30}{s/\sqrt{n}} > t_{1-\alpha,n-1}\Big|\mu = \mu_a\Big).$$

The power of the test is the probability that the T statistic lies in the rejection region. If this normalised mean was greater than the $t_{1-\alpha,n-1}$ quantile, then we reject the null hypothesis. Note that this quantity is calculated not under the assumption of the null hypothesis but rather calculated under the assumption of the alternative hypothesis. Note that this function depends on the specific value of μ_a and we notice that, as $\mu \to 30$, power approaches α.

2.1.5. *Calculating Power.* Assume that n is large, thus the central limit theorem holds - allowing us to use standard normal calculations rather than T calculations-, and we know σ, then

$$\begin{aligned}
1 - \beta &= \mathbb{P}\Big(\frac{\bar{X} - 30}{\sigma/\sqrt{n}} > z_{1-\alpha}\Big|\mu = \mu_a\Big) \\
&= \mathbb{P}\Big(\frac{\bar{X} - \mu_a + \mu_a - 30}{\sigma/\sqrt{n}}\Big|\mu = \mu_a\Big) \\
&= \mathbb{P}\Big(\frac{\bar{X} - \mu_a}{\sigma/\sqrt{n}} > z_{1-\alpha} - \frac{\mu_a - 30}{\sigma/\sqrt{n}}\Big|\mu = \mu_a\Big) \\
&= \mathbb{P}\Big(Z > z_{1-\alpha} - \frac{\mu_a - 30}{\sigma/\sqrt{n}}\Big|\mu = \mu_a\Big)
\end{aligned}$$

In the first line we have our test statistic $\frac{\bar{X}-30}{\sigma/\sqrt{n}}$, which under the null hypothesis $H_0 : \mu = 30$ is a Z-statistic, but we are going to calculate it under the alternative hypothesis. Hence, our rejection region will be for the normalised mean values larger than a standard normal quantile. Then $1 - \beta$ is the probability we reject a false null hypothesis, it's the probability that the statistic is larger than the quantile (the cutoff value), given that μ is in fact μ_a. Since we are considering the alternative hypothesis and not the null hypothesis, this quantity -$\frac{\bar{X}-30}{\sigma/\sqrt{n}}$- is no longer a Z-statistic but it's normal should the data be iid Gaussian.

In line 2, we convert it into a Z-statistic by adding and subtracting μ_a. In line 3 we take the correctly normalised mean $\frac{\bar{X}-\mu_a}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$ which is in fact a Z-statistic (under the alternative hypothesis), and we calculate the probability that said quantity is larger than $z_{1-\alpha} - \frac{\mu_a-30}{\sigma/\sqrt{n}}$. In line 4, we calculate the probability of Z being larger than $z_{1-\alpha} - \frac{\mu_a-30}{\sigma/\sqrt{n}}$, which can be easily computed provided σ is known and a μ_a value has been selected.

Example

Statement: Suppose we wanted to detect an increase in mean RDI of at least 2 events/hours above 30. Assume normality and setting a type I error rate of 5% and that the sample in question will have a standard deviation of 4, what would be the power if we took a sample size of 16?
Now, $Z_{1-\alpha} = 1.645$ and $\frac{\mu_a-30}{\sigma/\sqrt{n}} = 2$. Then $\mathbb{P}(Z > 1.645 - 2) = \mathbb{P}(Z > -0.355) = 64\%$
So under these settings, the probability of detecting an alternative of two events above the hypothesised value per hour is 64%. Note that this is only a bound for all values above 32%, the power gets larger as the alternative goes away from 30 events per hour.

Example of a Sample Size Calculation

Suppose we have a power we want to achieve for a particular value of the alternative, what sample size would we need to achieve it? What n would be required to get a power of 80%?
We want to compute the following expression and solve for n

$$0.8 = \mathbb{P}\Big(Z > z_{1-\alpha} - \frac{\mu_a - 30}{\sigma/\sqrt{n}}\Big| \mu = \mu_a\Big)$$

where the Z-statistic is in fact normalised under the alternative hypothesis. We set $z_{1-\alpha} - \frac{\mu_a - 30}{\sigma/\sqrt{n}} = z_{0.20}$, the 20th quantile of the standard normal distribution and solve for n. This guarantees a 80% or higher power, so μ_a is typically the smallest effect which can be reasonably detected. The calculation for $H_a : \mu < \mu_0$ is similar.

In general, for $H_a : \mu \neq \mu_0$ we calculate the one sided power using $\alpha/2$ (this is only approximately right since it excludes the probability of getting a large TS in the opposite direction of the truth).

In conclusion

- Power goes up as α gets larger,
- Power of a one sided test is greater than the power of the associated two sided test,
- Power goes up as μ_1 gets further away from μ_0 and power goes up as n goes up.

Regarding the first item, if we're requiring less evidence to reject the null hypothesis, we're bound to detect more alternative hypotheses.

2.1.6. *T Tests and Monte Carlo.*

Power for the T test. : Consider calculating power for a Gossett's T test for our example. The power is

$$\begin{aligned}
\mathbb{P}\Big(\frac{\bar{X} - 30}{S/\sqrt{n}} > t_{1-\alpha,n-1}\Big|\mu = \mu_a\Big) &= \mathbb{P}\Big(\sqrt{n}(\bar{X} - 30) > t_{1-\alpha,n-1}S\Big|\mu = \mu_a\Big) \\
&= \mathbb{P}\Big(\frac{\sqrt{n}(\bar{X} - 30)}{\sigma} > t_{1-\alpha,n-1}\frac{S}{\sigma}\Big|\mu = \mu_a\Big) \\
&= \mathbb{P}\Big(\frac{\sqrt{n}(\bar{X} - \mu_a)}{\sigma} + \frac{\sqrt{n}(\mu_a - 30)}{\sigma} > \frac{t_{1-\alpha,n-1}}{\sqrt{n-1}} \times \sqrt{\frac{(n-1)S^2}{\sigma^2}}\Big|\mu = \mu_a\Big) \\
&= \mathbb{P}\Big(Z + \frac{\sqrt{n}(\mu_a - 30)}{\sigma} > \frac{t_{1-\alpha,n-1}}{\sqrt{n-1}}\sqrt{\chi^2_{n-1}}\Big|\mu = \mu_a\Big),
\end{aligned}$$

where $Z \sim \mathcal{N}(0, 1)$ is the independent standard normal and where χ^2_{n-1} is the chi-squared random variable with $n - 1$ degrees of freedom. In regards to the first expression, the power is the probability that our test statistic lies in the rejection region and it's performed under the alternative hypothesis and, in regards to the second-to-last expression, if the data are iid Gaussian $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ and can be computed with the Monte-Carlo method, by simulating pairs of Z random variables and χ^2 random variables and evaluating this inequality for every pair. Note that, in general, μ_a per sé needs not to be specified, what we need to know is how different is μ_a from μ_0 in standard deviation units. This quantity, $\frac{\mu_a - \mu_0}{\sigma}$, is called an effect size is a unit free quantity. For example, a Monte Carlo simulation can be performed to calculate the last expression for the RDI example with the following R code

```r
no_sim <- 100000 # number of simulation to perform
n_dof <- 16 # number of degrees of freedom
sigma <- 4 # variance
mu0 <- 30 # RDI mean under the null hypothesis
mua <- 32 # RDI mean under the alternative hypothesis
z <- rnorm(no_sim) # rnorm is the R function that simulates random variables
    having a specified normal distribution
```

```
chisq <- rchisq(no_sim, df = n_dof - 1) # chi squared distribution
t_qt <- qt(.95, n_dof-1) # 95th quantile for the Gossett's T distribution
mean(z + sqrt(n_dof)*(mua-mu0)/sigma >
    t_qt/sqrt(n_dof-1)*sqrt(chisq))

In[1]: [1] 0.60517
In[2]: function (x, ...)
In[3]: UseMethod("mean")
In[4]: <bytecode: 0x56366e096b50>
In[5]: <environment: namespace:base>
```

In the previous code snippet, we have a 100000 pairs of normal and chi squared random variables, generated with the `rnorm` and `rchisq` R-functions. The `mean(...)` function returns a vector of ones every time the LHS is bigger than the RHS and zeros every time the LHS is smaller than the RHS and computes the mean of said vector entries. The accuracy of this computation is upto $\mathcal{O}\left(\frac{1}{n^2}\right)$.

2.1.7. *Two Sample Tests - Matched Data I.* When comparing two groups, first and foremost we want to determine whether observations are paired or not[6]. When dealing with a single set of paired data, one strategy is to take the difference between the paired observation and do a single one-sample t test of $H_0 : \mu_d = 0$ versus $H_0 : \mu_d \neq 0$ (or one of the other two alternatives). The desired test statistic

$$\frac{\bar{X}_d - \mu_{d0}}{S_d/\sqrt{n}}$$

where μ_{d0} is the value under the null hypothesis (typically 0), where n is the number of pairs of observations (and not the total number of observations themselves). This is called the **ordinary paired two-group t-test**.

2.1.8. *Two Sample Tests - Matched Data II.* Let's consider an example of a paired T-test.

Example

Consider Exam 1 and Exam 2 grades from a previous class. Is there any evidence that the second exam was easier or harder than the first?

Clearly, the data are paired since it's the same students measured twice. This question can be answered by examining if the mean of exam one is different from the mean of exam two.

In general, giver our data, we can compute this with the following R code

```
diff <- test2 - test1  #pair difference
n <- sum(!is.na(diff)) #number of subjects: 49
mean(diff) #mean of the pair difference: 2.88
sd(diff) #standard deviation of the pair difference: 7.61
testStat <- sqrt(n) * (mean(diff) - 0)/sd(diff) #test Statistic: 2.65

#we get the p-value by multiplying the probability of getting a test statistic
# as large or larger than 2.65 for a Gossett's T distribution with n-1 dof by two, since it's a two sided test
2 * pt(abs(testStat), n-1, lower.tail = FALSE) # Since we're working with 48 dof, there's little diference from calculating a pnorm or a pt
```

[6] For example, consider a medical trial of some medication. In one case we randomize that treatment to one group of test subjects and randomize a placebo to the other group of test subjects. This is not paired as the groups are distinct. An instance where the data are paired is, for example, administering the medication and the placebo in some random order for every test subject. Now the data is paired since some test subjects received both the treatment and the control.

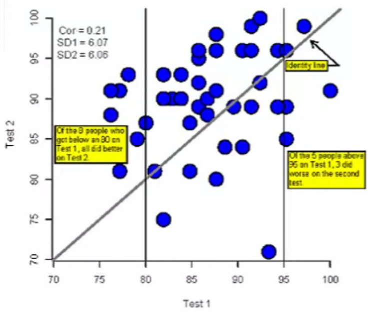


Figure 7. A matched data plot for the Exam 1 results v. Exam 2 results.

```

------


## or using the R function
t.test(diff)
```

In our case we get that there appears to be a difference between the means of exam 1 and exam 2. [7].

2.1.9. *Two Sample Tests - Regression to the Mean.* Francis Galton was the first to recognise that for matched data, high initial observations tended to be associated with lower second observations and low initial observations observations tended to be associated with higher second observations. For example, sons of very tall fathers tend to be a little shorter and fathers of very tall sons tend to be shorter. Another example would be that second exams for those who scored very high on a first exam tend to be a little lower.

In order to explain this phenomena, consider plot 7. Imagine if the tests were completely random and the students were iid draws from said distribution, so the higher observations of the exam were random observations. So the probability of a second observation being that high is quite low, since it's more likely to be in the center of the distribution. Conversely, a very low test (something that had a very low probability of occurring given that it's already low), the probability of a second test being that low is small. And so if the pairs of observations are exactly noise, then you'll have a lot of regression in the mean. This is one extreme of complete variation and, for a given student, there's no trend between exam 1' score and exam 2' score.

Consider the other extreme, let's imagine that the test was a perfect adjudicator of student's abilities, that it is a perfectly calibrated instrument and that there is no noise. Then the student should ideally get exactly the same score on both exams. At which point, there'd be no variation around an identity line of Exam 2 v. Exam 1. This is an extreme of no variation where there is a 100% correlation. Every other practical case lies somewhere in between those two extremes. ¡

For example, of the eight people who got below 80 en the test, all did better on Test 2 and of the five people who got above a 95 test, three did worse on the second test.

To investigate more, we normalise both scales (so that their empirical means are both 0 and their empirical standard deviations are 1). When asking questions about the paired data, we're asking questions about shifts in the means, information which we've gotten rid off. If there was no regression to the mean, the data would scatter about an identity line. The best fitting line goes through the average and has slope

$$\mathrm{Cor}(Test1, Test2)\frac{SD(Test2)}{SD(Test1)}$$

and passes through the point

[7] Graphically this can be represented with a so called "Mean difference plot", first presented by Tukey and by J. Martin Bland & Douglas G. Altman.

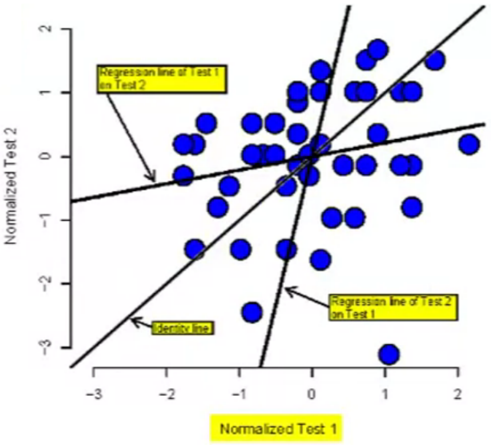


Figure 8. Normalised Test 2 scores v. Normalised Test 1 scores and the best fitting lines. In this plot we have the normalised Test 2 in the y-axis, the normalised Test 1 in the x-axis and the slope of the best regression line of Test 1 on Test 2 is the correlation between the two (0.21). In the middle we plot the identity line whilst the best regression line of Test 2 on Test 1 the slope is the inverse of the previous slope 1/0.2. Note that all three lines pass through the (0,0).

$$(\text{mean}(Test1), \text{mean}(Test2))$$

Since we normalised the data, the best fitting line passes through (0,0) and has a slope of $\text{Cor}(Test1, Test2) < 1$, since renormalizations don't affect the correlation. This will be shrunk toward a horizontal line, telling us our expected normalised test score for Test 2 is $\text{Cor}(Test1, Test2)$ times the normalised Test 1 score. This line appropriately adjusts for regression to the mean for Test 2 conditioning on Test 1. We could similarly do the same for Test 1 conditioning on Test 2(in this case the best fitting line will have slope $\text{Cor}(Test1, Test2)^{-1}$ if we plot with Test 1 on the Test 1 on the horizontal axis). The latter line will be shrunk toward a vertical line, the identity line will fall between the two. This can be better understood in plot 8. In said plot we have the normalised Test 2 in the y-axis, the normalised Test 1 in the x-axis and the slope of the best regression line of Test 1 on Test 2 is the correlation between the two (0.21). In the middle we plot the identity line whilst the best regression line of Test 2 on Test 1 the slope is the inverse of the previous slope 1/0.2. Note that all three lines pass through the (0,0).

The line to be used when predicting test 2 scores from test 1 scores is a very flat (suggesting there's little correlation between the two tests, and this amount of noise also suggests there's a fair amount of regression to the mean) line whilst the line used to predict test 1 scores from test 2 scores is a very vertical line (because our correlation was quite low, suggesting a high amount of noise). Should the points collapse around an identity line, this suggests there's very little regression.

In conclusion:

- An ideal examiner would have little difference between the identity line and the fitted regression line.
- The more unrelated the two exam scores are the more pronounced the regression to the mean is.

2.1.10. *Two Sample Tests - Two Independent Groups.* The extension to two independent groups is straightforward. Let $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$ (or one of the other two alternatives). Assuming a common error variance we have that the following statistic

$$\frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{n_X+n_Y-2},$$

under the null hypothesis $\bar{X} - \bar{Y}$ has a hypothesised mean of 0 and if the data are iid Gaussian. In the limit of a large sample size, the previous statistic follows a normal distribution. If the assumption of a common error variance is questionable then

$$\frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim \mathcal{N}(0, 1) \text{ if } n_x, n_y \to \infty.$$

where S_p is the pooled variance. Furthermore, this statistic follows an approximate (only approximately, since the variances are not the same) Student's T distribution if $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$. The approximate degrees of freedom are

$$\frac{(S_x^2/n_x + S_y^2/n_y)^2}{(S_x^2/n_x)^2/(n_x - 1) + (S_y^2/n_y)^2/(n_y - 1)}$$

where S_x and S_y are the variances within group 1 and group 2 respectively.

Note that the connection between hypothesis testing and confidence intervals still holds: for example, if zero is in our independent group T interval, we will fail to reject the independent group T test for equal means and viceversa, if we construct a confidence interval by finding all of the hypothesised values for differences of means for which we fail to reject the null hypothesis we wind up with the appropriate t confidence interval.

If we want to test equality of means, in general it's incorrect constructing separate confidence intervals, one CI for group I and another CI for group II, and seeing if those CIs overlap. This procedure only works if the CIs don't overlap and we reject, this is an accurate statement. But the confidence intervals can overlap and the correctly constructed test statistic would still reject, seemingly leading to a contradiction. This procedure has lower power than the correctly implemented test.

Exam

Suppose that instead of having repeated data on two consecutive exams, students were randomised to two teaching modules and took the same exam. We treat the data as independent group data. We might obtain data like the following

Group	N	Mean Exam	SD Exam
Module 1	50	86.9	6.07
Module 2	50	89.8	6.06

The pooled standard deviation is 6.065 and the test statistic is

$$\frac{89.8 - 86.9}{6.065\sqrt{\frac{1}{50} + \frac{1}{50}}}$$

Suppose you have equal numbers of observations for two groups, X and Y. If the data are truly matched, then the standard error of the difference is estimating

$$\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{n} - 2\frac{\text{Cov}(X, Y)}{n}}.$$

If we ignore the matching by setting $\mathrm{Cov}(X,Y)=0$, the standard error of the difference is estimating is estimating

$$\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{n}}.$$

Since, generally, matched data is positively correlated, by ignoring the matching, we are unnecessarily augmenting the standard error.

2.2. **Week 2.** The score statistic is a specific two sample binomial test that will serve as motivation for creating a confidence interval

2.2.1. *Two sample Binomial Tests - Score Statistic.* Consider the following example

Example

Consider a randomized trial where 40 subjects were randomized (20 each) to two drugs with the same active ingredient but different expedients. Consider counting the number of subjects with side effects for each drug. The gathered data is
—— Side Effects —— None —— Total
Drug A —— 11 —— 9 —— 20
Drug B —— 16 —— 15 —— 20
Total —— 16 —— 14 —— 40

At first sight, there appears to be a higher propensity for side effects from Drug A than from Drug B. We'd like to do a test of whether or not the propensity of side effects is the same within the two drugs.

Let's start with score tests. Consider testing the null hypothesis $H_0 : p = p_0$ for a binomial proportion. The **score test statistic** is constructed in the same way as a Z-test. Consider a single binomial proportion, only looking at drug A and considering if drug A has a specific population proportion of side effects

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \underset{n\to\infty}{\sim} \mathcal{N}(0,1).$$

Note that $\sqrt{p_0(1-p_0)/n}$ is the standard error of the binomial distribution and we're using the null hypothesis[8]. This test performs better than the Wald test

$$\frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}}.$$

Both of the previous statistics are compared to the quantiles of the standard normal distribution, the upper $\frac{\alpha}{2}$th quantile for a two-sided test or the upper αth quantile for an upper single-sided test where the alternative is $H_a : p > p_0$ and the αth quantile for an upper single-sided test where the alternative is $H_a : p < p_0$.

Inverting the Wald test yields the Wald interval, namely those values of p_0 for which we'd fail to reject the null hypothesis:

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

Inverting the Score test yields the Score interval

$$\hat{p}\Big(\frac{n}{n + z^2_{1-\frac{\alpha}{2}}}\Big) + \frac{1}{2}\Big(\frac{z^2_{1-\frac{\alpha}{2}}}{n + z^2_{1-\frac{\alpha}{2}}}\Big) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n + z^2_{1-\frac{\alpha}{2}}}\Big[\hat{p}(1-\hat{p})\frac{n}{n + z^2_{1-\frac{\alpha}{2}2}} + \frac{1}{4}\frac{z^2_{1-\frac{\alpha}{2}}}{n + z^2_{1-\frac{\alpha}{2}}}\Big]}$$

Note that $\frac{n}{n+z^2_{1-\frac{\alpha}{2}}}$ and $\frac{z^2_{1-\frac{\alpha}{2}}}{n+z^2_{1-\frac{\alpha}{2}}}$ add up to one, it's a point in the two-dimensional simplex. If $n \to \infty$ the first term dominates and $\hat{p}$ dominates, conversely if n is small the second term dominates. Plugging $z_{1-\frac{\alpha}{2}} = 2$ yields the Agresti-Coull interval. As $n \to \infty$ this interval gets progressively similar to the Wald interval. This test takes $\hat{p}$ and shrinks it down to $\frac{1}{2}$ which is desired since for $p > \frac{1}{2}$, the binomial distribution gets progressively asymmetric.

[8] Should we use $\hat{p}$, we'd be constructing the estimated standard error, akin to the calculations of a confidence interval.

Example continued

In our previous example consider testing whether or not Drug A's percentage of subjects with side effects is greater than 10%. Then we'd like to test $H_0 : p_A = .1$ versus $H_a : p_A > .1$. Let $\hat{p} = 11/20 = .55$. Then our test statistic is

$$\frac{.55 - .1}{\sqrt{.1 \times .9/20}} = 6.7$$

If we are performing a two-sided 95th Z-test, the quantile of interest is 1.645 and if we're performing a one sided 95th Z-test the quantie of interest is 1.96. In any case, $6.7 > z_{95\text{th}}$, so we reject Our P-value, the probability of getting a z bigger than 6.7 is nearly around 0 (we're about six standard deviations from zero for a standard normal carries a very low probability since three standard deviations already covers the majority of the distribution). For a two-sided test, we'd double this P-value.

We're postulating the number of side effects out of 20 is a binomial trial. Implicit in this idea is the postulate of iid data, every person is an independent and identically distribution, drawn from a population. We use these ideas to construct a super population that has a prevalence of side effects of p_A. So our iid model is giving us an idea of a population proportion and we're testing relative to that proportion.

2.2.2. *Two sample Binomial Tests - Exact Tests.* The previous method relies on the central limit theorem and having a large-enough sample size. It's possible to perform and exact binomial test. Consider calculating an exact P-value. What's the probability, under the null hypothesis, of getting evidence as extreme or more extreme than the one we obtained (ie. the probability of getting more than 11 people with side effects)?

$$P(X_A \geq 11) = \sum_{x=11}^{20} \binom{20}{x} (.1)^x (.9)^{20-x} \approx 0.$$

This calculation, the probability of getting more than 11 people with side effects out of 20, is done under the null hypothesis $H_0 : p_0 = 10\%$. This is the probability of getting evidence as or more extreme - in favour of the alternative - with the probability being calculated under the null hypothesis. In conclusion, the probability of getting more than 11 people of 20 under the null hypothesis of $p = .1$ is approximately 0.

This can be computed with the following R-routine

```
pbinom(10, 20, .1, lower.tail = FALSE)

In [1]: 7.088606e-07
---------------

binom.test(11, 20, .1, alternative="greater")

In [1]: Exact binomial test
In [2]: data:  11 and 20
In [3]: number of successes = 11, number of trials = 20, p-value = 7.089e-07
In [4]: alternative hypothesis: true probability of success is greater than 0.1
In [5]: 95 percent confidence interval:
In [6]:  0.3469314 1.0000000
In [7]: sample estimates:
In [8]: probability of success
In [9]:                     0.55
```

Note that in R, `lower.tail = FALSE` calculates the probability of strictly greater than 10 (starts at 11 and ends at 20) and, `lower.tail = TRUE` calculates the probability of strictly less or equal than 10.

This test, unlike the asymptotic ones, guarantees the Type I error rate is less than the desired level (usually this desired level is 5%), sometimes it's much less. Inverting the exact binomial test yields an exact binomial interval for the true proportion. This interval, called the Clopper-Pearson interval, has coverage greater than 95%, though can be very conservative. For two sided tests, calculate the two one-sided P-values and double the smaller.

2.2.3. *Two sample Binomial Tests - Comparing 2 Binomial Proportions.* Now we want to compare two proportions. Consider now testing whether the proportion of side effects is the same in the two groups or different. Let $X \sim \text{Binomial}(n_1, p_1)$, let $\hat{p}_1 = \frac{X}{n_1}$, let $Y \sim \text{Binomial}(n_2, p_2)$ and let $\hat{p}_2 = \frac{Y}{n_2}$. We also standardise notation as follows

$$\begin{array}{lll} n_{11} = X, & n_{12} = n_1 - X & n_1, = n_{1+}, \\ n_{21} = Y, & n_{22} = n_2 - Y & n_2, = n_{2+}, \\ n_{+1}, n_{+2} & 1 & \end{array}$$

Consider testing $H_0 : p_1 = p_2$ versus one of the three usual alternatives. The score test statistic for this null hypothesis is

$$TS = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

where $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions for each binomial distribution, where the variance of the difference of the random variables is $\hat{p}(1-\hat{p})$ for $\hat{p} = \frac{X+Y}{n_1+n_2}$, the estimate of the common proportion under the null hypothesis. We need an estimated version of this variance, so as to compare it to a normal quantile. That is, we need a P-value to plug in. We can plug $\hat{p}$ if under the null hypothesis the sample proportions are identical, the group A is made up of iid draws from Bernoulli 1 and the group B is made up of iid draws from Bernoulli 2, but since they are common we have $n_1 + n_2$ Bernoulli draws and our proportion would simply be the total number of events. This statistic is normally distributed for large n_1, n_2.

The interval doesn't have a closed form inverse for creating a confidence interval and must be numerically calculated. An alternate interval inverts the Wald test:

$$TS = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}.$$

This Wald test doesn't use the fact that, under the null hypothesis, the proportions are equal, having separate $\hat{p}_1$ and $\hat{p}_2$.

The resulting confidence interval is

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

As in the one sample case, the Wald interval and Wald test perform very poorly, relative to the score interval and score test. For testing, we always use the score test whilst for interval, inverting the score test is hard and not offered in standard software.

A simple fix is the Agresti-Caffo interval, where we add one failure and one success to each group, which is obtained by calculating $\tilde{p}_1 = \frac{X+1}{n_1+2}$, $\tilde{n}_1 = n_1 + 2$, $\tilde{p}_2 = \frac{Y+1}{n_2+2}$, $\tilde{n}_2 = n_2 + 2$. Using these we can simply construct the Wald interval. This interval does not approximate the score interval but does perform better than the Wald interval.

Example

Test whether or not the proportion of side effects is the same for the two drugs: $\hat{p}_A = .55$, $\hat{p}_B = 5/20 = .25$, $\hat{p} = 16/40 = .4$. The test statistic is

$$\frac{.55 - .25}{\sqrt{.4 \times .6 \times (1/20 + 1/20}} = 1.61.$$

We fail to reject H_0 at .05 level by comparing to 1.96. The P-value is $P(|Z| \geq 1.61) = .11$

Likelihood analysis requires the use of so-called profile likelihoods or some other technique to reduce dimensions. Consider putting independent Beta(α_1, β_1) and Beta(α_2, β_2) with priors on p_1 and p_2, respectively. Then the posterior is likelihood times prior equals posterior. In our case the likelihood function for a Beta distribution is

$$L(\theta) = \frac{\Gamma(1+\theta)}{\Gamma(1)\Gamma(\theta)} \prod_{i=1}^{n} (1 - Y_i)^{\theta - 1},$$

which simplifies to

$$L(\theta) = \theta (1 - Y_i)^{n(\theta - 1)}.$$

Thus, the posterior is

$$\pi(p_1, p_2) \propto p_1^{x + \alpha_1 - 1} (1 - p_1)^{n_1 + \beta_1 - 1} \times p_2^{y + \alpha_2 - 1} (1 - p_2)^{n_2 + \beta_2 - 1}.$$

Hence under this, potentially naive prior, the posterior for p_1 and p_2 are independent Beta's. The easiest way to explore this posterior is via Monte Carlo simulation.

2.2.4. *Relative Risks and Odds Ratios - Relative Measures.* There are many instances in which we prefer to talk about an odds ratio or a relative measure instead of an absolute measure. For example, for an event that's rare, some environmental effect which causes a fairly rare disease, where we're comparing a small proportion of people who contracted the disease (among the unexposed group) and a small proportion of people who contracted the disease among the exposed group. The absolute difference in rates is very small, the relative difference might be very large.

Motivation

Consider a randomized trial where 40 subjects were randomized (20 each) to two drugs with the same active ingredient but different expedients. Consider counting the number of subjects with side effects for each drug:

—— Side effects —— None —— Total

Drug A —— 11 —— 9 —— 20

Drug B —— 5 —— 15 —— 20

Total —— 16 —— 14 —— 40

We are interested in whether drug A has a statistically higher percentage of side effects than drug B, accounting for what would be expected by chance.

Let $X \sim$ Binomial(n_1, p_1), $Y \sim$ Binomial(n_2, p_2), let $\hat{p}_1 = \frac{X}{n_1}$ and let $\hat{p}_2 = \frac{Y}{n_2}$ are the estimators for the proportions of each binomial distribution.

2.2.5. *Relative Risks and Odds Ratios - The Relative Risk.* Now we are interested in relative changes, particularly useful when both proportions are small. The **relative risk** is defined as p_1/p_2. The natural estimator for the relative risk is

$$\hat{RR} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{X/n_1}{Y/n_2},$$

Naturally, this relative risk has an issue if Y has zero counts. The standard error for $\log \hat{RR}$, much more useful for constructing confidence intervals, is

$$\hat{SE}_{\log \hat{RR}} = \sqrt{\frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}}.$$

One can construct a confidence interval for the log of the relative risk by adding and subtracting a standard normal quantile and then we multiply said standard normal quantile by the standard error of the log relative risk. By exponentiating the resulting interval we get an interval for the RR.

Alternatively we can take the ratio of odds instead of the ratio of probabilities. So the population odds ration is

$$\frac{\text{Odds of SE Drug A}}{\text{Odds of SE Drug B}} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2 2(1-p_1)}.$$

and we'll compare this number against 1. If it's bigger than one, it's going to suggest a greater propensity of side A and if it's less than 1, it's going to suggest a smaller propensity for side effects of drug A. The sample odds ratio simply plugs in the estimates for p_1 and p_2, working out to be

$$\hat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{n_{11}n_{22}}{n_{12}n_{21}},$$

the so-called cross product ratio. The log of the estimated odds ratio has better asymptotic Gaussian behaviour, then the standard error for $\log \hat{OR}$ is

$$\hat{SE}_{\log \hat{OR}} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

Thus, in order to construct a confidence interval, we calculate the odds ratio, we log it and calculate the standard error of the log odds ratio, we add or subtract a standard normal quantile (say 1.96 for a 95% confidence interval); thus obtaining a confidence interval for the log odds ratio. By exponentiating the end point, we obtain an interval for the OR. Thre are some problems of course if $n_1 = 0$ or $n_2 = 0$.

Notice that the sample and true odds ratios do not change if we transpose the rows and the columns. For both the OR and RR, taking the logs helps with adherence to the error rate. Of course the interval for the $\log RR$ or $\log OR$ is obtained by taking

$$\text{Estimate} \pm z_{1-\frac{\alpha}{2}} SE_{\text{Estimate}},$$

and, as usual, exponentiating yields an interval for the OR and RR in the natural scale. Though logging helps, these intervals still don't perform altogether that well because of the asymptotics of these statistics being ruled by the central limit theorem (if we plug in a 95th quantile, we don't necessarily get a 95th confidence interval).

2.2.6. *Relative Risks and Odds Ratios - The Odds Ratio.* Let's consider the previous example

Exampled continued

For the relative risk, $\hat{p}_A = 11/20 = .55$ and $\hat{p}_B = 5/20 = .25$. The estimated relative risk is $\hat{RR}_{A/B} = .55/.55 = 2.2$ and the estimated standard error for the logs is $\hat{SE}_{\log \hat{RR}_{A/B}} = \sqrt{\frac{1-.55}{.55\times 20} + \frac{1-.25}{.25\times 20}} = .44$. Then the interval for the log RR is $\log(2.2) \pm 1.96 \times .44 = [-.07, 1.65]$ and the interval for the RR is $[.93, 5.21]$.

For the odds ratio we have that the estimated odds ratio is $\hat{OR}_{A/B} = \frac{11\times 15}{9\times 5} = 3.67$. Also $\hat{SE}_{\log \hat{OR}_{A/B}} = \sqrt{\frac{1}{11} + \frac{1}{9} + \frac{1}{5} + \frac{1}{15}} = .68$, with the interval for the $\log OR$ being $\log(3.67) \pm 1.96 \times .68 = [-.04, 2.64]$ and the interval for the OR being $[.96, 14.01]$.

For the estimated risk difference $\hat{RD}_{A-B} = \hat{p}_A - \hat{p}_B = .55 - .25 = .30$, with an estimated standard error of $\hat{SE}_{\hat{RD}_{A-B}} = \sqrt{\frac{.55\times .45}{20} + \frac{.25\times .75}{20}} = .15$ and the confidence interval being $.30 \pm 1.96 \times .15 = [.15, .45]$.

2.2.7. *Delta Method.* We're interested in obtaining the standard error and test statistics for these methods using the Delta method.

Recall $X \sim \text{Binomial}(n_1, p_1)$, $Y \sim \text{Binomial}(n_2, p_2)$ where we have

- $\hat{RD} = \hat{p}_1 + \hat{p}_2$, $\hat{SE}_{\hat{RD}} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$,
- $\hat{RR} = \frac{\hat{p}_1}{\hat{p}_2}$, $\hat{SE}_{\log \hat{RR}} = \sqrt{\frac{1-\hat{p}_1}{\hat{p}_1 n_1} + \frac{1-\hat{p}_2}{\hat{p}_2 n_2}}$,
- $\hat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$, $\hat{SE}_{\log \hat{OR}} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$

with the confidence interval being $CI = \text{Estimate} \pm z_{1-\frac{\alpha}{2}} SE_{\text{Estimate}}$.

The delta method can be used to obtain large sample standard errors for instances where we're not longer dealing with differences of averages but, for example, with the logs of some desired quantities. Let $\hat{\theta}$ be an estimator for the quantity θ with the estimator given by

$$\frac{\hat{\theta} - \theta}{\hat{SE}_\theta} \to \mathcal{N}(0, 1).$$

Formally, the delta method states that for a sufficiently smooth function $f : Dom(f) \subset \mathbb{R} \to Im(f) \subset \mathbb{R}$ then

$$\frac{f(\hat{\theta}) - f(\theta)}{f'(\hat{\theta})\hat{SE}_\theta} \to \mathcal{N}(0, 1),$$

ie. the asymptotic mean of $f(\hat{\theta}$ is $f(\theta$.

2.2.8. *Delta Method and Derivation.* Let $\theta = p_1$ and the estimator $\hat{\theta} = \hat{p}_1 = \frac{X}{n_1}$, the estimated standard error is $\hat{SE}_{\hat{\theta}} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}$. Let $f(x) = \log(x)$, then by the central limit theorem

$$\frac{\hat{\theta} - \theta}{\hat{SE}_{\hat{\theta}}} \to \mathcal{N}(0, 1),$$

since $\hat{\theta}$ is a simple average and dividing it by it's standard error. Then

$$\begin{aligned}\hat{SE}_{\log \hat{p}_1} &= f'(\hat{\theta})\hat{SE}_{\hat{\theta}} \\ &= \frac{1}{\hat{p}_1}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} = \sqrt{\frac{1-\hat{p}_1}{\hat{p}_1 n_1}},\end{aligned}$$

and

$$\frac{\log \hat{p}_1 - \log p_1}{\sqrt{\frac{1-\hat{p}_1}{\hat{p}_1 n_1}}} \to \mathcal{N}(0,1).$$

If we want a confidence interval for $\log p_1$, we can add and subtract to a standard normal quantile (say 1.96 for a 95% interval) times this standard error.

The asymptotic standard error for the log relative risk is

$$\operatorname{Var}(\log \hat{RR}) = \operatorname{Var}\Big(\log \frac{\hat{p}_1}{\hat{p}_1}\Big), \tag{1}$$

$$= \operatorname{Var}(\log \hat{p}_1) + \operatorname{Var}(\log \hat{p}_2) \tag{2}$$

$$\approx \frac{1-\hat{p}_1}{\hat{p}_1 n_1} + \frac{1-\hat{p}_2}{\hat{p}_2 n_2}, \hat{SE}_{\log \hat{RR}} = \sqrt{\frac{1-\hat{p}_1}{\hat{p}_1 n_1} + \frac{1-\hat{p}_2}{\hat{p}_2 n_2}} \tag{3}$$

where the last line follows from the multivariate delta method. This approximation requires large sample sizes. Thus the delta method gives both a standard error as well as a variance calculation and an asymptotic normality, giving a way to create confidence intervals.

Heuristically, let's assume we have a large enough sample size. If $\hat{\theta}$ is close to θ then

$$\frac{f(\hat{\theta}) - f(\theta)}{\hat{\theta} - \theta} \approx f'(\hat{\theta}),$$

so

$$\frac{f(\hat{\theta}) - f(\theta)}{f'(\hat{\theta})} \approx \hat{\theta} - \theta,$$

therefore

$$\frac{f(\hat{\theta}) - f(\theta)}{f'(\hat{\theta})\hat{SE}_{\hat{\theta}}} \approx \frac{\hat{\theta} - \theta}{\hat{SE}_{\hat{\theta}}} \to \mathcal{N}(0,1).$$

2.3. Week 3.

2.3.1. *Fisher's Exact Test.* Fisher's exact test is exact because it guarantees the α rate, regardless of the sample size. When performing an asymptotic test, using a nominal type I error rate of say 5%, and we calculate a 95% confidence interval for the risk difference and declare the differences in the proportions as being significant, if the confidence interval for the difference doesn't include zero, that's a nice useful valid testing procedure. However this doesn't guarantee a 5% error rate, only asymptotically as the sample size goes to infinity. Fisher's exact test, in contrast, guarantees a 5% limited provided the iid assumptions are met for each of the two groups.

Example

Let's consider a chemical toxicant and 10 mice, and we treat five with the toxicant and five with the control, as follows
—— Tumor —— None —— Total
Treated —— 4 —— 1 —— 5
Control —— 2 —— 3 —— 5
Total —— 6 —— 4
where the last line corresponds to the number of tumors for the treated versus control. Let's assume we have two binomials and we wish to test equality of the proportions by fixing the margins, 5 treated and 5 controlled for each group. We're modelling the probability that a random mouse from this population of treated mice has a tumour as being p_1 and similarly the probability p_2 that a random mouse from this population of controlled mice having a tumour.

Let the null hypothesis be $H_0 : p_1 = p_2 = p$, where p is the common proportion. We're not able to use the Z-test nor the χ^2-test since the sample size is small and since we don't have a specific value for p, necessary for both tests. Under the null hypothesis every permutation is equally likely. Imagine the treatment and control status were randomised. Then, if the null hypothesis is true, it would be exchangeable for any mouse - whether or not it got a tumor or was from the treated group or the control group -, Let the observed data be, for example,

- Treatment: T(reated) T T T T C C C C C,
- Tumor: T(umor) T T T T N(o tumor) T T N N N

then the permuted data could be

- Treatment: T C C T C T T C T C,
- Tumor: T T T T N T T N N N.

Note that the total number of treated and the total number of controlled remained fixed whilst the total number of tumor and the total number of non tumors also remained fixed. This seems like a reasonable null distribution to investigate, this is that the treatment and control statistics are exchangeable relative to tumor status, so we'll look at some test statistic relative to this distribution. The consequence of this is, every time we permute treatment and control labels, if we were to reform the two by two table it would have the same margins (five on the row margins and six and four on the column margins).

Fisher's exact test uses this null distribution to test the hypothesis that $p_1 = p_2$, by explicitly using the idea of randomisation.

2.3.2. *Hyper-Geometric Distribution.* In the previous section we were basing our analysis on conditions over the data (fixed margins and fixed number of tumors and non-tumors) as well as the randomisation process, under the hypothesis the randomisation is irrelevant (whether any given mouse received treatment or control is irrelevant as long as the margins remain fixed). Let X be the number of tumors for the treated group and Y is the number of tumors for the control group. Let $H_0 : p_1 = p_2 = p$ be the null hypothesis. Under the null hypothesis we have

$$X \sim \text{Binom}(n_1, p), \qquad Y \sim \text{Binom}(n_2, p), \qquad X + Y \sim \text{Binom}(n_1 + n_2, p),$$

where the last identity follows precisely because, under the null hypothesis, X and Y are a number of iid Bernoulli draws. These assumptions being made regardless of sample size. We're desiring to construct a probability distribution which doesn't depend with the unknown parameter, p. Then

$$\mathbb{P}\Big(X = x \Big| X + Y = z\Big) = \frac{\binom{n_1}{x}\binom{n_2}{z-x}}{\binom{n_1+n_2}{z}},$$

which is the hypergeometric probability mass function, which doesn't depend on p (the so-called conditioning on a sufficient statistic). In effect, we can prove the previous result:

$$\begin{aligned}
\mathbb{P}\Big(X = x\Big) &= \binom{n_1}{x} p^x (1-p)^{n_1 - x} \\
\mathbb{P}\Big(Y = z - x\Big) &= \binom{n_2}{z - x} p^{z-x} (1-p)^{n_2 - z + x} \\
\mathbb{P}\Big(X + Y = z\Big) &= \binom{n_1 + n_2}{z} p^z (1-p)^{n_1 n_2 + x}
\end{aligned}$$

The first line is simply the probability of a binomial random variable $X \sim \text{Binom}(n_1, p)$ taking the value x. The second line is the probability of Y and $z - x$, where $z - x \in \mathbb{Z}_{[0, n_2]}$, and the last line is the probability of $X + Y$ equal to z. Then

$$\begin{aligned}
\mathbb{P}\Big(X = x \Big| X + Y = z\Big) &= \frac{\mathbb{P}\Big(X = x, X + Y = z\Big)}{\mathbb{P}\Big(X + Y = z\Big)} \\
&= \frac{\mathbb{P}\Big(X = x, Y = z - x\Big)}{\mathbb{P}\Big(X + Y = z\Big)} \\
&= \frac{\mathbb{P}(X = x)\mathbb{P}(Y = z - x)}{\mathbb{P}(X + Y = z)}
\end{aligned}$$

where it readily follows that

$$\mathbb{P}\Big(X = x \Big| X + Y = z\Big) = \frac{\binom{n_1}{x}\binom{n_2}{z-x}}{\binom{n_1+n_2}{z}}.$$

2.3.3. *Fisher's Exact Test in Practice and Monte Carlo.* Let's consider an experiment where we have more tumors for the treated group than for the control group and we wish to calculate an exact P-value using the conditional distribution. The conditional distribution fixes both the row and the column totals. The calculations are permutation-invariant, yielding the same results independently if the rows or columns are fixed. The hypergeometric distribution, derived as a conditional distribution, is identical to the permutation distribution, which randomly permutes treatment and control labels by stringing the data out as the full data set and not as just a two-by-two table.

All one-sided versions of Fisher's exact test yield the same inference. For two-sided tests, not all test statistics are equal.

Consider the alternative hypothesis $H_a : p_1 > p_2$. The P-value requires tables as extreme or more extreme, under the alternative hypothesis, than the one observed. Note we are fixing both the row and the column totals. Let the observed table be

$$\text{Table 1} = \left|\begin{array}{c|c|c} 5 & 4 & 1 \\ 5 & 2 & 3 \\ 10 \text{ total} & 6 & 4 \end{array}\right|$$

The only more extreme table in favour of the alternative is

$$\text{Table 1} = \left|\begin{array}{c|c|c} 5 & 5 & 0 \\ 5 & 1 & 4 \\ 10 \text{ total} & 6 & 4 \end{array}\right|$$

where instead of only four mice from the treated group getting the tumor, all five of them got the tumor and the rest of the cells are fixed since the margins are fixed.

Then the probabilities are computed using the hypergeometric distribution

$$\mathbb{P}(\text{Table 1}) = \mathbb{P}(X = 4 | X + Y = 6) = \frac{\binom{5}{4}\binom{5}{2}}{\binom{10}{6}} = 0.238$$
$$\mathbb{P}(\text{Table 2}) = \mathbb{P}(X = 5 | X + Y = 6) = \frac{\binom{5}{5}\binom{5}{1}}{\binom{10}{6}} = 0.024$$

Thus the P-value is is the sum $0.238 + 0.024 = 0.262$. Thus the only way to construct a 5% confidence interval would be with the last dataset, the most extreme table. This is a consequence of the exact testing. These test guarantee at most a 5% error rate, not exactly 5% since the data is discreet and there are only so many probabilities availeable to the P-value. In this example, the only way to reject would be by getting the most extreme table. This calculation can be computed with the following R-routine

```
dat <- matrix(c(4,1,2,3), 2)
fisher.test(dat, alternative="greater")

#### output

In [1]: Fishers Exact Test for Count Data
In [2]: data:  dat
In [3]: p-value = 0.2619
In [4]: alternative  hypothesis: true odds ratio is greater than 1
In [5]: 95 percent confidence interval:
In [6]:  0.3152217        Inf
In [7]: sample estimates:
In [8]: odds ratio
In [9]:   4.918388
```

The simplest way to obtain a two-sided P-value can be is by doubling the smaller of the two one-sided P-value (so as to not obtain a larger than one P-value). The other way for creating a two-sided test statistic, a test statistic that measures whether a table is as or more extreme than the observed table. One example is using the chi-squared test statistic, then calculating the hypergeometric probability for every two by two table satisfying the margins, adding up the probabilities associated with those tables, with the chi-squared statistic, that are bigger (ie. more in favour of the alternative) than the observed table. In this setting, the problem is there's no uniformly most powerful statistic, thus every election for a test statistic results in a power trade-off. [9]

[9] Fisher's statistic were the hypergeometric probabilities themselves. In his test, we'd calculate the probabilities for all tables that satisfied the margins and for every table with a hyper geometric probability smaller than the observed hypergeometric

The discreteness of the problem usually dictates a large P-value for small sample sizes. In the previous example, the second-most extreme table we could possibly obtain had a low probability, winding up with a 26% P-value. Note that this exact method doesn't distinguish between rows or columns.

The common value for P under the null hypothesis is called a nuisance parameter. Condition on the total number of successes for both data groups, $X+Y$, eliminates this nuisance parameter p. Also, Fisher's exact test guarantees the type I error rate as a bound and not exactly.

Another form of obtaining a P-value is with the alternative exact unconditional test can be found as

$$\sup_P \mathbb{P}\Big(\frac{X}{n_1} > \frac{Y}{n_2}; p\Big).$$

Fisher's exact test with Monte Carlo

Let the observed table have $X = 4$ and be

- T T T T T C C C C C

- T T T T N T T N N N

One thing we could do is to permute the first row:

- T C T T C C C T T T

- T T T T N T T N N N

A Monte-Carlo simulation can be perform and output a hypergeometric P-value. For example, a simulated table has $X = 3$ and we would repeat this calculation a great many number of times and calculate the proportion of tables for which the simulated $X \geq 4$ (ie. evidence as or more extreme, in favour of the alternative, than the one observed). This proportion is a Monte Carlo estimate for Fisher's exact P-value

2.3.4. *Chi Squared Testing.* We're interested in chi-squared testing for contingency tables and the most classic contingency table test is testing independence. We'll relate that to testing independence of several proportions, generalisations to higher order contingency tables, and Monte-Carlo variations to get the exact tested independence as well as analysing a special kind of independence test, a fit-test which is used to test if the data arrive from a particular distribution.

An alternative approach to testing equality of proportions uses the chi-squared statistic

$$\sum \frac{(\text{Observed} - \text{Expect})^2}{\text{Expected}} \sim \chi^2_1,$$

where "Observed" are the observed counts, "Expected" are the expected counts under the null hypothesis and the sum is performed for all four cells. The Chi-squared statistic is exactly the square of the difference in proportions Score statistic. It's a weighted distance between observed values and expected theoretical values. Note this test is impervious to directionality.

probability, would had it's probability summed up to the observed hypergeometric probability, thus obtaining the P-value. The logic for performing this kind of test is that if something arose out of not the null distribution but rather from an alternative distribution, then those tables would have a low probability under the null hypothesis, thus the hypergeometric probabilities are a useful test statistic.

Example Chi-squared testing

Let's consider a medical trial where X and Y are two equally effective treatments and we desire to know if one adds more side effects than the other.

Trt	Side Effects	None	Total
X	44	56	100
Y	77	43	120
	121	99	220

Here we assigned 100 test subjects to X and 120 test subjects to Y, and our model consists in treating the number of people with side effects out of the total as if they were binomial random variables. Let p_1 and p_2 be the rates of side-effects for X and Y, respectively. The null hypothesis is $H_0 : p_1 = p_2$.

The χ^2-statistic is $\sum \frac{(O-E)^2}{E}$ where

- $O_{11} = 44$, $E_{11} = \frac{121}{220} \times 100 = 55$,
- $O_{21} = 77$, $E_{21} = \frac{121}{220} \times 120 = 66$,
- $O_{12} = 56$, $E_{12} = \frac{99}{220} \times 100 = 45$,
- $O_{22} = 43$, $E_{22} = \frac{199}{220} \times 100 = 54$,

Note that if the rate of side effects was the same for the two groups, for the two treatments, then our estimate for the common proportion would have to be

$$\frac{\text{Total side effects for X} + \text{Total side effects for Y}}{\text{Total number of test subjects}} = (44 + 77)/220,$$

and we'd use this value as our best estimate for the overall proportion of side effects, regardless of treatment since under the null hypothesis the proportions are assumed equal. Then the number of people we'd expect to see with side effects, for the first group, is $\frac{121}{220} \times 100 = 55$ and $\frac{121}{220} \times 120 = 66$ for the second group. Then for the first group we'd expect to see 1-(total side effects)/(total of test subjects) $= (1 - (44 + 77)/220) \times 100 = 45$ test subjects without side-effects and 1-(total side effects)/(total of test subjects) $= (1 - (44 + 77)/220) \times 120 = 54$ for the second group. Note that these calculations are performed under the null hypothesis using our best estimate for the null hypothesis, $\frac{\text{Total side effects for X+Total side effects for Y}}{\text{Total number of test subjects}}$, where the margins 100 and 120 are fixed by design. Should these expected counts differ substantially from the observed count, then this would be evidence against the null hypothesis.

$$\chi^2 = \frac{(44-55)^2}{55} + \frac{(77-66)^2}{66} + \frac{(56-45)^2}{45} + \frac{(43-54)^2}{54} = 8.96.$$

Then we'd compare it to a χ^2 with one degree of freedom to reject for larger values (we favor the alternative hypothesis of the further away from the expected value counts). So the bigger the test statistic is, we're going to favour the alternative, rejecting for large values. This result suggests there's a difference in the rate of side effects between the treatments, but we don't know about the directionality (ie. which one is larger).

The χ^2-value for a one degree of freedom can be computed using the following R-routine

```
pchisq(8.96, 1, lower.tail = FALSE)
In [1]: 0.002
```

We can readily perform a Chi-squared testing with the following R-routine

```
dat <- matrix(c(44, 77, 56, 43), 2)
chisq.test(dat)
chisq.test(dat, correct = FALSE)

In [1]: Pearsons Chi-squared test with Yates continuity correction

In [2]: data:  dat
In [3]: X-squared = 8.1667, df = 1, p-value = 0.004267


In [4]: Pearsons Chi-squared test

In [5]: data:  dat
In [6]: X-squared = 8.963, df = 1, p-value = 0.002755
```

We don't get exactly the same test-statistic than the one calculated previously since the chi-squared approximation is an asymptotic approximation. But since the counts are discrete, it's possible to improve the chi-squared approximation by the means of continuity correction.

In conclusion, we reject if the statistic is too large. The alternative is always two sided, since we're testing if the proportions are equal or not and we do not divide the type I error rate by two, even though we're performing a two-sided test. A small χ^2 implies little difference between the observed values and those expected under H_0. The χ^2 statistic and approach generalises to other kinds of tests and larger contingency tables.

An alternative computational form for the χ^2 statistic is

$$\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{+1}n_{+2}n_{1+}n_{2+}}.$$

where

$n_{11} = X$	$n_{12} = n_1 - X$	$n_1 = n_{1+}$
$n_{21} = Y$	$n_{22} = n_2 - Y$	$n_1 = n_{2+}$
n_{+1}	n_{+2}	n

Notice that the statistic

$$\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{+1}n_{+2}n_{1+}n_{2+}}.$$

doesn't change if you transpose the rows and the columns of the table. Surprisingly, the χ^2 statistic can be used in the following situations:

- the rows are fixed (binomial),
- the columns are fixed (binomial),
- the total sample size if fixed (multinomial),
- none are fixed (Poisson).

For a given set of data, any of these assumptions results in the same value for the statistic.

2.3.5. *Testing Independence.* Now we'll use the χ^2 test to analyse independence. Consider the following example

Example 1 of Independence Testing

We're interested in studying Maternal age vs. birthweight. The experiment's sample size is of 400 test subjects but the margins are not fixed ie. the cross-sectional sample is such that only the total sample size is fixed.

		Birthweight	
Mat. Age	$< 2500g$	$\geq 2500g$	Total
$< 20y$	20	80	100
$\geq 20y$	30	270	300
Total	50	350	400

We'll treat the cell counts as if they were four dimensional multinomial counts with $N = 400$. We would like to know is if the variable birth weight is independent of maternal age. Let the null hypothesis be H_0 : MA is independent of BW and let the alternative hypothesis be H_a : MA is not independent of BW.

Our estimate for the young maternal age can be obtained as

$$\mathbb{P}(\text{MA} < 20) = \frac{100}{400} = .25,$$

the estimated marginal probability of low birth weight is

$$\mathbb{P}(\text{BW} < 2500) = \frac{50}{400} = .125.$$

Thus, under the null hypothesis, the estimated cell probabilities are readily found to be

- younger and low birth: $\mathbb{P}(\text{MA} < 20 \text{ and BW} < 2500) = \frac{100}{400} \times \frac{50}{400}$,
- younger and high weight: $\mathbb{P}(\text{MA} < 20 \text{ and BW} \geq 2500) = \frac{100}{400} \times \frac{300}{400}$,
- older and low birth: $\mathbb{P}(\text{MA} \geq 20 \text{ and BW} < 2500) = \frac{300}{400} \times \frac{50}{400}$,
- older and high birth: $\mathbb{P}(\text{MA} \geq 20 \text{ and BW} \geq 2500) = \frac{300}{400} \times \frac{350}{400}$.

Note that this calculation is performed under the null hypothesis since the events are independent. Therefore the expected counts are

- $E_{11} = \frac{100}{400} \times \frac{50}{400} \times 400 = 12.5$,
- $E_{12} = \frac{100}{400} \times \frac{350}{400} \times 400 = 87.5$,
- $E_{21} = \frac{300}{400} \times \frac{50}{400} \times 400 = 37.5$,
- $E_{22} = \frac{300}{400} \times \frac{350}{400} \times 400 = 262.5$.

Then the χ^2 is

$$\chi^2 = \frac{(20-12.5)^2}{12.5} + \frac{(80-87.5)^2}{87.5} + \frac{(30-37.5)^2}{37.5} + \frac{(270-262.5)^2}{262.5} = 6.86.$$

We can compare with the critical value `qchisq(.95,1) = 3.84` and calculate a P-value (the probability of getting a test statistic as large as 6.86 or larger) `pchisq(6.86,1, lower.tail = FALSE) = 0.009`.

The answer obtained in this test is the same as the one obtained in the test proportions, obtaining the same test statistic and P-value, but with a different interpretation (ie. if there was randomisation of the rows).

Let's consider another example:

Example 2 of Independence Testing

We're interested in cross-classifying profession bi alcohol use by studying 300 clergymen, 250 educators, 300 Executives and 350 retailers for a total of 1200 test subjects. Does alcohol use differ by occupation?

Group	High Al. Use	Low Al. Use	Total
Clergy	32	268	300
Educator	51	199	250
Executives	67	233	300
Retailers	83	267	350
Total	233	967	1200

Our interest lies in testing whether or not the proportion of high alcohol use is the same in the four occupations:
$H_0 : p_1 = p_2 = p_3 = p_4 = p$,
and the alternative hypothesis is

$$H_a : \text{at least two of the} p_j \text{are unequal}$$

Then, under the null hypothesis (all proffesions have the same rate of high alcohol use), our obvious estimate for the common proportion of high alcohol use is

$$p = \frac{\sum_{i=1}^{4} \text{Test Subjects with High Al. Use}}{4} = \frac{233}{1200}.$$

Then the observed counts and expected values are

- $O_{11} = 32$, $E_{11} = 300 \times \frac{233}{1200}$,
- $O_{12} = 268$, $E_{12} = 300 \times \frac{967}{1200}$,
- $\cdots$

Note that since the margins are fixed, the expected and observed counts must add up to the margins. Then the chi-squared statistic is

$$\sum \frac{(O-E)^2}{E} = 20.59$$

with the number of degrees of freedom given by

$$df = (\text{Rows} - 1) \times (\text{Columns} - 1) = 3.$$

The P-value can be found with the following R-routine: `pchisq(20.59, 3, lower.tail = FALSE)` ≈ 0.

This means that some of them are indeed different.

2.3.6. *Generalisation.* Let's now analyse a generalisation to the chi-squared test for word distributions in the following example:

Example 1 of Generalised Chi-squared Independence Testing

We're interested in whether the word distributions of these words is equivalent across the three books and we've sampled so many words from each of the books. Let the data be:

Wood	Book 1	Book 2	Book 3	Total
a	147	186	101	434
an	25	26	11	62
this	32	39	15	86
that	94	105	37	236
with	59	74	28	161
without	18	10	10	38
Total	375	440	202	1017

Our null hypothesis is the H_0 : the probabilities of each word are the same for every book, and our alternative hypothesis is H_a : At least two are different. We have a multinomial distribution for every column and we desire to test the proportions of these multinomial distributions.
Then under our null hypothesis, the proportions for a particular word across the three books are equal, so our estimated proportion for "a" would be

$$p_a = \frac{\sum_{i=1}^{4} \text{"a"-counts for each book}}{4} = \frac{434}{1017}.$$

The observed counts and expected counts, under the null hypothesis, are

- $O_{11} = 147$, $E_{11} = 375 \times \frac{434}{1017}$,
- $O_{12} = 186$, $E_{12} = 440 \times \frac{434}{1017}$,
- $O_{21} = 25$, $E_{12} = 375 \times \frac{62}{1017}$,
- $\cdots$

where O_{ij} is the number of times the ith-word appeared in jth-book and where

$$E_{ij} = \text{Total of words in the } i\text{-book} \times p_j,$$

this is, the ij-th expected count is the product of the total number of all words in the ith-book times the estimated proportion for the jth word. Then, the χ^2 statistic is

$$\sum \frac{(O-E)^2}{E} = 12.27$$

with the degrees of freedom being

$$df = (6-1)(3-1) = 10.$$

Then, by comparing it to the relevant cut-off gives the result to the experiment.

Example 2 of Generalised Chi-squared Independence Testing

We're interested in rating couples based sexual intercourse rates. The sample size is 91 couples and we're interested in cross-classifying them. Let the data be:

Husband	N	F	V	A	Tot
N	7	7	2	3	19
F	2	8	3	7	20
V	1	5	4	9	19
A	2	8	9	14	33
	12	28	18	33	91

where N: never, F: fairly often, V: very often, A: almost always. Our null hypothesis is H_0 : H and W ratings are independent, our alternative hypothesis is H_a : not independent.
Under the null hypothesis, for example,

$$\mathbb{P}(H = N \text{ and } W = A) = \mathbb{P}(H = N)\mathbb{P}(W = A).$$

Our relevant statistic is

$$\chi^2 = \sum \frac{(O - E)^2}{E}.$$

For example, our observed counts (under the null hypothesis) is $O_{11} = 7$ whilst the expected counts are $E_{11} = 91 \times \frac{19}{91} \times \frac{12}{91} = 2.51$. In general, $E_{ij} = \frac{n_{i+}n_{+j}}{n}$ with the number of degrees of freedom being $df = (\text{Rows} - 1)(\text{Columns} - 1)$.

We can compute this test with the following R routine

```
x <- matrix(c(7,7,2,3,
              2,8,3,7,
              1,5,4,9,
              2,8,9,14),4)
chisq.test(x)

In [1]: Pearson's Chi-squared test

In [2]: data:  x
In [3]: X-squared = 16.955, df = 9, p-value = 0.04942

In [4]: Warning message:
In [5]: In chisq.test(x) : Chi-squared approximation may be incorrect
```

Which gives

$$\sum \frac{(O - E)^2}{E} = 16.96,$$

with $df = 9$ degrees of freedom with a P-value of .049. Note that the chi-squared test are asymptotic tests, validated by the central limit theorem. The cell counts might be too small to use a large sample approximation.

These equal distribution and independence test yield the same results if

- the row totals are fixed,
- or the column total are fixed,
- or the total sample size is fixed,
- or none are fixed.

Note that mathematically equivalent results applied in different setting may result in wildly different interpretations.

We can use a Monte-Carlo approximation to calculate an exact P-value for contingency tables

Exact Permutation Test

Imagine if we got the individual data points as well as the contingency table. Let the raw data be

W: NNNNNNNNFFFFFFFVVAAANNFFFFFFF$\cdots$.
H: NNNNNNNNNNNNNNNNNNNNNNNFFFFFF$\cdots$.

If the Wife's rating and the Husband's rating were independent, then the matching of the pairs of ratings would be irrelevant (we can map the correct husband to the correct wife or not, according to the null hypothesis this shouldn't matter at all). Thus we can permute either the W or H rows. Recalculate the contingency table we would have the same margins (the same number of husbands answering N, the same number of wives answering N and so on), we can then calculate the χ^2 statistic for each permutation and calculate the percentage of times it is larger than the observed value is the exact P-value by using the R function: `chisq.test(x, simulate.p.value = TRUE`.

2.3.7. *Goodness of Fit Testing.* Let's now talk about goodness of fit testing. We're interested in testing R's uniform random number generator. Let the data obtained from a thousand simulations be

Example 1 of Goodness of Fit Testing

	[0,.25)	[.25,.5)	[.5,.75)	[.75,1)	Total
Count	254	235	267	244	1000
TP	.25	.25	.25	.25	1

The question to answer is how many uniforms are expected to see between zero and .25, .25 and .5, .5 and .75 and between .75 and 1.0? Therefore our the null hypothesis is $H_0 : p_1 = p_2 = p_3 = p_4 = .25$ and the alternative hypothesis is H_a : any $p_i \neq p_i^0$, where p_i^0 is the ith hypothesised value.
We can calculate the expected counts using the assumed probability density function. Then

- $O_1 = 254$, $E_1 = 1000 \times .25 = 250$,
- $O_2 = 235$, $E_2 = 1000 \times .25 = 250$,
- $O_3 = 267$, $E_3 = 1000 \times .25 = 250$,
- $O_4 = 244$, $E_1 = 1000 \times .25 = 250$,

and

$$\sum \frac{(O-E)^2}{E} = 2.264,$$

with $df = \text{Cells} - 1 = 3$, which in turn gives a P-value of .52. Note this test only test the uniformness of the random number generator only.

Let's now consider another example.

Example 2 of Goodness of Fit Testing

In Mendel's P plant experiment, we are give the following data for phenotype.

	Yellow	Green	Total
Observed	6022	2001	8023
TP	.75	.25	1
Expected	6017.25	2005.75	8023

where our null hypothesis is $H_0 : p_1 = .75, p_2 = .25$ with our test statistic being

$$\sum \frac{(O-E)^2}{E} = \frac{(6022-6017.25)^2}{6017.25} + \frac{(2001-2005.75)^2}{2005.75} = .015.$$

He have a P-value of .90 for a single degree of freedom. Now note that Fisher originally combined several of Mendel's tables such that

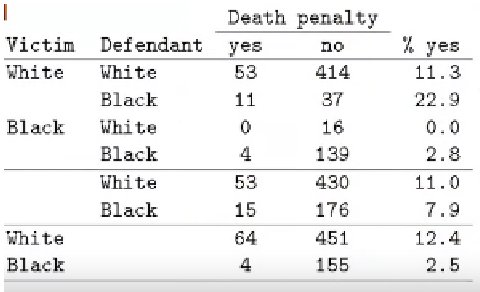
$$\sum \chi^2_{\nu_i} \sim \chi^2_{\sum \nu_i}$$

The statistic was 42 with 84 degrees of freedom for a total P-value of .99996, founding that the data fit Mendel's model too well.

Let's summarise: For a goodness of fit testing, we are testing whether or not the observed counts correspond to equal theoretical value by using the chi-squared test statistic:

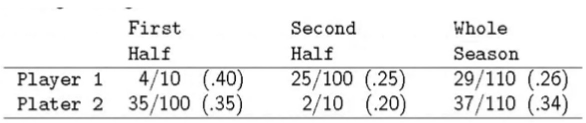
$$\sum \frac{(O-E)^2}{E}.$$

This is an asymptotic test, since this test statistic follows a χ^2 distribution for large n with the degrees of freedom being the number of cells minus 1. There are others test, such as the Kolmogorov-Smirnov test, which do not require discretization.



Victim	Defendant	Death penalty yes	no	% yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
	White	53	430	11.0
	Black	15	176	7.9
White		64	451	12.4
Black		4	155	2.5

FIGURE 9. Data Groups



	First Half	Second Half	Whole Season
Player 1	4/10 (.40)	25/100 (.25)	29/110 (.26)
Plater 2	35/100 (.35)	2/10 (.20)	37/110 (.34)

FIGURE 10. Caption

2.4. **Week 4.**

2.4.1. *Simpson's Paradox.* We're interested in a cross classification of defendants from criminal trials by the race of the victim in murder trials and their relation to a death penalty verdict. Consider the data given in 9.

By close inspection of the data we see that white defendants receive the death penalty a fewer percentage of the time, 11% compared to 22%, for both white and black victims. Zero of the white defendants received the death penalty whilst 2.8% for the black victims.

But, if we disregard the race of the victim, it turns out that white defendants receive the death penalty a greater percentage of the time, 11% compared to 7.9%. And if we look the race of the victim, disregarding the race of the defendant, actually in the instance where the victim was white, the defendant received the death penalty a higher percentage of the time, 12.4% compared to a 2.5%.

In this setting, what would be the conclusion about death penalty verdicts vs. defendants and victim's race? A priori, we get two totally different and contradictory answers, depending if we're analysing the raw data itself or the marginal probabilities. Which one is the right answer?

Let's analyse the data:

- Marginally, white defendants received the death penalty a greater percentage of the time than black defendants.
- Across white and black victims, black defendant's received the death penalty a greater percentage of the time than white defendants.

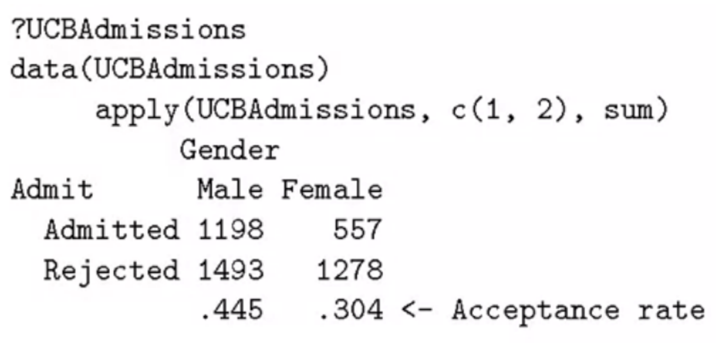
Simpson's paradox refers to the fact that marginal and conditional associations can be opposing. The death penalty was enacted more often for the murder of a white victim than a black victim. Whites tend to kill whites (demographically), hence the larger marginal association. Thus, Simpson's paradox states that two variables can change when factoring a third variable[10].

2.4.2. *Simpson's Paradox, more examples.* Consider now the following example. We're interested in comparing two player's batting averages. Let the data be

Player 1 has a better batting average than Player 2 in both the first and second half of the season, yet has a worse batting average overall. Player 1 had a very good batting average when they had relatively few bats and a modest batting average when they had lots of bats and vice-versa for player two.

A famous example of Simpson's paradox is present in UC Berkeley Admissions data. We're given a certain dataset and and the total number of admitted and rejected applications for males and females, as it can be seen in 11.

[10] For further research, see Larry Wasserman's post "The normal deviate" on Simpson's Paradox.



```
?UCBAdmissions
data(UCBAdmissions)
     apply(UCBAdmissions, c(1, 2), sum)
          Gender
Admit      Male Female
  Admitted 1198    557
  Rejected 1493   1278
           .445   .304 <- Acceptance rate
```

Figure 11. Berkeley Admissions data

It turns out the male acceptance rate was higher than the females' acceptance rate, disregarding everything else. But if we study the acceptance rates by departments, it turns out that in almost every department, males had a lower acceptance rate than females. There are different gender acceptance rates by department.

Mathematically, Simpson's paradox is not paradoxical. Let $a, \cdots, h \in \mathbb{Z}$, then

$$\text{if } a/b < c/d \text{ and } e/f < g/h \Rightarrow (a+e)/(b+f) > (c+g)/(d+h)$$

In statistical terms, it says that the apparent relationship between two random variables can change in the light or absence of a third random variable. It's only when we conflate the probabilistic statements and the evidence associated with probabilistic statements, vis a vie the data, with the causal statements, that so-called paradoxes may arise. In general, it's a non-trivial problem to determine how much conditioning is enough on a given dataset so as to extract reliable conclusion. This problem of confounding is treated by the Causal inference sub-discipline.

2.4.3. *Weighting and Confounding.* Variable that are correlated with the explanatory and response variables can distort the estimated effect. In our previous example, victim's race was clearly correlated with the defendant's race and the death penalty verdict. With a confounder present, we're not able to distinguish an event being causally related with race in the death penalty experiment versus something that has a statistical association with race and death penalty verdict. We're interested in events or random variables which have a statistical association with the explanatory and response variables, where there is a plausible causal connection between them.

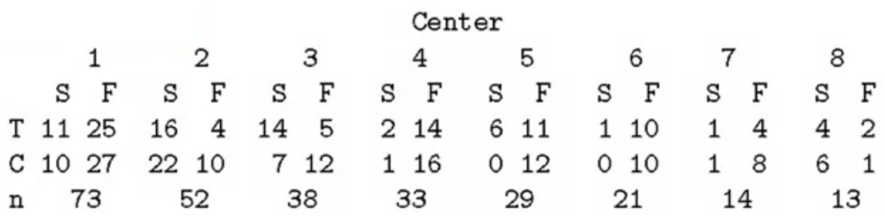
Let's assume we have a single confounder. How do we deal with it? One way is with regression, whilst another strategy in categorical data analysis to adjust for confounding variables is to stratify by the confounder and then combine the strata-specific estimates. This requires appropriately weighting the strata-specific estimates. Note that unnecessary stratification reduces precision.

Suppose we have two unbiased scales, one with a variance of 1lb and the other one with variance of 9lbs. Confronted with weights from both scales, would you give both measurements equal creedance?

Suppose that we weigh a certain object with the first scale. This is represented by a random variable which we assume to be normal, $X_1 \sim \mathcal{N}(\mu, \sigma_1^2)$. Similarly, weighing this object in the second scales gives another random variable which we assume to be normal as well, $X_2 \sim \mathcal{N}(\mu, \sigma_2^2)$. Note that both σ_1 and σ_2 are known. We'd like to estimate μ, then the log-likelihood for μ is

$$-\frac{(x_1-\mu)^2}{2\sigma_1^2} - \frac{(x_2-\mu)^2}{2\sigma_2^2} + \cdots,$$

where the dots stand for terms which don't depend on μ. We take the derivative with respect to μ and we set it to zero:



	Center															
	1		2		3		4		5		6		7		8	
	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F
T	11	25	16	4	14	5	2	14	6	11	1	10	1	4	4	2
C	10	27	22	10	7	12	1	16	0	12	0	10	1	8	6	1
n	73		52		38		33		29		21		14		13	

Figure 12. Dataset for a medical trial in the context of a CMH test.

$$-\frac{(x_1-\mu)}{\sigma_1^2}-\frac{(x_2-\mu)}{\sigma_2^2}=0$$
$$\Rightarrow \frac{x_1r_1+x_2r_2}{r_1+r_2}=x_1p_1+x_2(1-p),$$

where $r_i=\frac{1}{\sigma_i^2}$ and $p=\frac{r_1}{r_1+r_2}$. Note that if X_1 has very low variance, then its term dominates the estimate for μ. The general principle thus being: instead of averaging over several unbiased estimates, take a single average weighted according to the inverse variances. In our example, $\sigma_1^2=1$, $\sigma_2^2=9$ so $p=.9$.

2.4.4. *C Mantel Haenszel Estimator and Test.* Consider the death penalty statistical trial. We have two two-by-two tables, were we are given the data of the defendant's race and whether or nor they got a death penalty verdict and then we stratified that by a third random variable, the victim's race. Thus, let n_{ijk} be the ijth-entry of the kth-table. In other words, $k=1$ is the first victim's race, $k=2$ is the second victim's race and so on, then i and j would index defendant's race and whether or not the person got the death penalty. The kth-sample odds ratio (main diagonal elements divided by the off-diagonal elements) is given by

$$\hat{\theta}_k=\frac{n_{11k}n_{22k}}{n_{12k}n_{21k}},$$

The Mantel Haenszel estimator is the weighted average of these strata-specific estimates and is given by

$$\hat{\theta}=\frac{\sum_k r_k\hat{\theta}_k}{\sum_k r_k} \text{ where the weights are } r_k=\frac{n_{12k}n_{21k}}{n_{++k}},$$

this is, in essence, a simplicial convex combination of the strata-specific odds ratios. The weights are in essence inverse variances of hypergeometric distributions. The estimator simplifies to[11]

$$\hat{\theta}_{MH}=\frac{\sum_k \frac{n_{11k}n_{22k}}{n_{++k}}}{\sum_k \frac{n_{12k}n_{21k}}{n_{++k}}}.$$

Let's now consider the following example of a medication's medical trial with the dataset given in 12. Let T be the active drug, C the placebo, S stands for success and F for failure. We're interested in whatever policies and practices existed at the various centers at which the data was collected by the means of a center-stratification with 8 strata. Then we have eight odds ratio and we wish to analyse if the center, as a random variable, was a confound.

Let's first imagine if the center was specifically associated with the treatment application (some centers tended to apply the treatment more than others) and that the center was associated with the success of the treatment due to different policies. In order to gain insight, we stratify by center, obtain a strata-specific odds ratio and then factoring in the inverse variance of the odds ratios (since some centers have more patients than others). For example, center 1, with 73 test subjects, has more weight in the final result than center 7, which only has 14 test subjects; since the first center has a better, more precise, odds ratio. Therefore, the Mantel Haenszel estimate is

[11] The standard error can be found in Agresti (page 235) or Rosner (page 656)

$$\theta_{MH} = \frac{(11\times 27)/73 + (16\times 10)/25 + \cdots + (4\times 1)/13}{(10\times 25)/73 + (4\times 22)/25 + \cdots + (6\times 2)/13} = 2.13,$$

also $\log\theta_{MH} = .758$ and $\hat{SE}_{\log\theta_{MH}} = .303$.

Another rationale for a stratified-estimate, is that we're under the impression that the center is at some level, a random effect, and is a modifier for the treatment. The centers studied are a random draw from the population of centers, and since we're not interested in a given center's specific (ie. we don't care if treatment worked at center number 1 or not), we care only if it worked overall. Therefore a CMH test is an average over a certain number of random draws from a given population of centers and quantifies it's effect on the treatment. This is called common odds ratio across centers.

Now, we're interested in testing if the common odds ratios are equal to 1. Let θ_i be the ith-strata-specific odds ratio, then let $H_0 : \theta_1 = \cdots = \theta_k = 1$ versus the alternative $H_a : \theta_1 = \cdots = \theta_k \neq 1$. The CMH test does apply to the other alternatives but it works best for the previous alternative hypothesis. This test is, in essence, the same as the test for conditional independence of response and exposure given the stratifying variable. In the Cochran Mantel Haenszel test is executed by conditioning on the rows (similarly as Fisher's exact test) and columns for each of the k contingency tables, resulting in k hypergeometric distributions and leaving only the n_{11k} cells free. Under the conditioning and under the null hypothesis,

$$\mathbb{E}(n_{11k}) = \frac{n_{1+k}n_{+1k}}{n_{++k}} \text{ and } \mathbb{V}\text{ar}(n_{11k}) = \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_{++k}^2(n_{++k}-1)}.$$

Then the CMH test statistic is

$$\frac{[\sum_k (n_{11k} - \mathbb{E}(n_{1+k}))^2]}{\sum_k \mathbb{V}\text{ar}(n_{11k})},$$

the sum of the deviations of the upper left-hand cells from their expected values, summed up. Contrast this to the chi-squared test, where they are first squared and then summed up. For many large sample sizes and under the null hypothesis, this test statistic is $\chi^2(1)$, regardless of how many table are being summed up. Remember that the CMH test is used when we're interested in analysing whether or not the odds ratio is one, given that it's common across all strata. An CMH can be performed in R using the following R routine

```
dat <- array(c(11, 10, 25, 27,  16, 22,  4, 10,
               14,  7,  5, 12,   2,  1, 14, 16,
                6,  0, 11, 12,   1,  0, 10, 10,
                1,  1,  4,  8,   4,  6,  2,  1),
             c(2,2,8))
mantelhaen.test(dat, correct=FALSE)

In [1]: Mantel-Haenszel chi-squared test without continuity correction

In [2]: data:  dat
In [3]: Mantel-Haenszel X-squared = 6.3841, df = 1, p-value = 0.01151
In [4]: alternative hypothesis: true common odds ratio is not equal to 1
In [5]: 95 percent confidence interval:
In [6]: 1.177590 3.869174
In [7]: sample estimates:
In [8]: common odds ratio
In [9]:         2.134549
```

The result for the CMH test statistic is 6.38 and we'll compare this to $\chi^2(1)$ value and we're going to reject for larger values. Since the P-value is 0.012 the CMH test suggest that the treatment and response are not conditionally independent given center.

It's possible to perform an analogous test in a random effects logit model that benefits from a complete model specification. It's also possible to test heterogeneity of the strata-specific odds radio. It's also possible to test if all the odds ratios are equal by the means of a Wolffe's tests. Note that exact test, which guarantee the type I error rate, are also possible in R by the means of the `exact = TRUE` command.

2.4.5. *Case Control Sampling.* In this section, we're going to treat case control methods, the instance where using using retrospective case control data and a so called rare disease assumption, we can estimate prospective odds ratios; and finally the exact inference for odds ratios.

Let's start by studying retrospective in reference sampling. We're interested in studying lung cancer cases and controls and to ascertain whether or not they were a smoker. There are two ways to collect this data: we could follow a given number of test subjects over time, some of them would smoke, and some of then wouldn't, and then we could quantify the amount of positive lung cancer diagnosis. This sampling method is inadequate and impossible for large sample sizes. Another form to obtain this data would be by finding a given number of positive lung cancer patients and find the same number of controls, at some level comparable to the first group. And the we'd retrospectively determine whether or not they were smokers. The margins for both groups are fixed (both are 709). In case-control studies, it's also the norm to find a control which very closely matches a specific case and repeat this procedure for all cases. Let the data be given by table 2.4.5.

Table 1 =

	Lung cancer		
Smoker	Cases	Controls	Total
Yes	688	650	1338
No	21	59	80
	709	709	1418

We're interested in ascertaining who is a smoker and whether or not the cases group had a greater proportion of smokers and to make prospective conclusions to this retrospective data. It's impossible to directly estimate $\mathbb{P}(\text{Case}|\text{Smoker})$, but it's possible to estimate $\mathbb{P}(\text{Smoker}|\text{Case})$ given the data. We can also estimate the odds ratios in particular, we're interested in the odds of being a case given that the test subject is a smoker relative to the odds of becoming a case given that the test subject is a non-smoker. But this is equivalent to finding the odds of being a smoker given that the test subject is a case relative to the odds of being a smoker given that the test subject is a case. This is

$$\frac{Odds(\text{Case}|\text{Smoker})}{Odds(\text{Case}|\text{Smoker}^c)} = \frac{Odds(\text{Smoker}|\text{case})}{Odds(\text{Smoker}|\text{case}^c)}.$$

Let C stand for Case, and S for smoker, then

$$\begin{aligned}\frac{Odds(\text{Case}|\text{Smoker})}{Odds(\text{Case}|\text{Smoker}^c)} &= \frac{\mathbb{P}(C|S)/\mathbb{P}(\bar{C}|S)}{\mathbb{P}(C|\bar{S})/\mathbb{P}(\bar{C}|\bar{S})} = \frac{\mathbb{P}(C,S)/\mathbb{P}(\bar{C},S)}{\mathbb{P}(C,\bar{S})/\mathbb{P}(\bar{C},\bar{S})} \\ &= \frac{\mathbb{P}(C,S)\mathbb{P}(\bar{C},\bar{S})}{\mathbb{P}(C,\bar{S})\mathbb{P}(\bar{C},S)},\end{aligned}$$

which is the probability of caseness and smokerness times by the probability of being neither a case nor a smoker divided by the product of the off-diagonal probabilities. Note that by exchanging C and S we get the same result since $\mathbb{P}(A,B) = \mathbb{P}(B,A)$.

Remember that the sample odds ratio is $\frac{n_{11}n_{22}}{n_{12}n_{21}}$, which remains unchanged if a row or column is multiplied by a constants, and is transpose-invariant (the odds ratio do not change by changing which factor is the outcome and which factor is the predictor).

Now, it turns out that the odds ratio is related to the relative risk. In effect

$$\begin{aligned}
OR &= \frac{\mathbb{P}(S|C)/\mathbb{P}(\bar{S}|C)}{\mathbb{P}(S|\bar{C})/\mathbb{P}(\bar{S}|\bar{C})} \\
&= \frac{\mathbb{P}(C|S)/\mathbb{P}(\bar{C}|S)}{\mathbb{P}(C|\bar{S})/\mathbb{P}(\bar{C}|\bar{S})} \\
&= \frac{\mathbb{P}(C|S)}{\mathbb{P}(C|\bar{S})}\frac{\mathbb{P}(\bar{C}|\bar{S})}{\mathbb{P}(\bar{C}|S)} \\
&= RR \times \frac{1-\mathbb{P}(C|\bar{S})}{1-\mathbb{P}(C|S)} \text{ where } RR = \frac{\mathbb{P}(C|S)}{\mathbb{P}(C|\bar{S})}
\end{aligned}$$

where in line 2 we reversed the odds ratio. Note that the odds ratio aproximates the relative risk if $\mathbb{P}(\bar{C}|\bar{S})$ and $\mathbb{P}(C|\bar{S})$ are nearly equal or very small. This is case when the case-ness is very small ie. regardless of whether a test subject smokes or not, the probability of lung cancer is very small. This is the case of the **rare disease assumption** which implies $OR \approx RR$. It's important to clarify that, for the rare disease assumption to be valid, it has to be rare among the exposed and non-exposed both, not rare overall.

Consider the following example

Example of Rare Disease Assumption

Let the data be given by the following table

Table 2 =

	Disease		
Exposure	Yes	No	Total
Yes	9	1	10
No	1	999	1000
	10	1000	1010

The estimated probability of the disease is $\hat{\mathbb{P}}(D) = \frac{10}{1010} \approx .01$, the odds ratios is $\hat{OR} = \frac{9\times999}{1\times1} = 8991$, $\hat{RR} = \frac{9/10}{1/1000} = 900$. Disease is rare overall but it's not rare among the exposed which explains why the odds ratio doesn't approximate the relative risk.

Let's do a recap:

- $OR = 1$ implies no association between the data,
- $OR > 1$ implies a positive association
- whilst $OR < 1$ implies a negative association.
- For retrospective case control studies, the odds ratio can be interpreted prospectively.
- For diseases rare among both the case and control groups, the odds ratio approximate the relative risk.
- The delta method gives the standard error for the logarithm of the odds ratio as

$$\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}$$

Going back to table 2.4.5 and the lung cancer example we have that the estimated odds ratio is $\hat{OR} = \frac{688\times59}{21\times650} = 3.0$, the estimated standard error for the logarithm of the estimated odds ratio is

$$\widehat{SE}_{\log \hat{OR}} = \sqrt{\frac{1}{688}+\frac{1}{650}+\frac{1}{21}+\frac{1}{59}} = .26.$$

The confidence interval at 95% confidence is $CI = \log(3.0) \pm 1.96 \times .26 = [.59, 1.61]$.

Note that the estimated odds of lung cancer for smokers is three times that of the odds for non-smokers with an interval of $[\exp(.59), \exp(1.61)] = [1.80, 5.00]$.

2.4.6. *Exact inference for the Odds Ratio.* Let X be the number of smokers for the cases and Y be the number of smokers for the controls. The margins are fixed and X and Y are random numbers modelled as binomial distributions. We desire to calculate an exact confidence interval for the odds ratio. In order to do that, first we need to eliminate the nuisance parameter.

Let $\text{logit}(p) = \log \frac{p}{1-p}$ be the log-odds. The differences in logits are the log-odds ratios $(\text{logit}(p_1) - \text{logit}(p_2) = \text{logit}(p_1/p_2)$.

$$\text{If } \text{logit}(\mathbb{P}(\text{Smoker}|\text{Case})) = \delta \text{ then } \mathbb{P}(\text{Smoker}|\text{Case}) = \frac{e^{\delta}}{1+e^{\delta}},$$

$$\text{If } \text{logit}(\mathbb{P}(\text{Smoker}|\text{Control})) = \delta + \theta \text{ then } \mathbb{P}(\text{Smoker}|\text{Control}) = \frac{e^{\delta+\theta}}{1+e^{\delta+\theta}},$$

where θ is the log-odds ratio and $\delta \in \mathbb{R}$ is the nuisance parameter, which we wish to eliminate. If $X \sim \text{Binom}\Big(n_1, p_1 = \frac{e^{\delta}}{1+e^{\delta}}\Big)$ and if $Y \sim \text{Binom}\Big(n_2, p_2 = \frac{e^{\delta+\theta}}{1+e^{\delta+\theta}}\Big)$, then

$$\begin{aligned}\mathbb{P}(X = x) &= \binom{n_1}{x}\Big(\frac{e^{\delta}}{1+e^{\delta}}\Big)^{x}\Big(\frac{1}{1+e^{\delta}}\Big)^{n_1-x} \\ &= \binom{n_1}{x} e^{x\delta}\Big(\frac{1}{1+e^{\delta}}\Big)^{n_1}.\end{aligned}$$

Then

$$\mathbb{P}(Y = z - x) = \binom{n_2}{z-x} e^{(z-x)\delta+(z-x)\theta}\Big(\frac{1}{1+e^{\delta+\theta}}\Big)^{n_2}.$$

Thus, since X and Y are not identically distributed we have

$$\mathbb{P}(X + Y = z) = \sum_{u} \mathbb{P}(X = u)\mathbb{P}(Y = z - u).$$

Finally, the probability that the random variable X takes a particular value x given that the sum of random variables $X + Y$ takes on a particular value z is

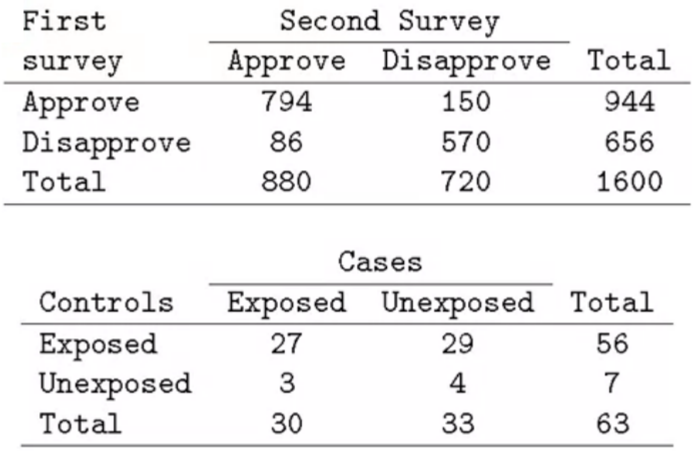
$$\begin{aligned}\mathbb{P}(X = x|X + Y = z; \theta) &= \frac{\mathbb{P}(X = x)\mathbb{P}(Y = z - x)}{\sum_u \mathbb{P}(X = u)\mathbb{P}(Y = z - u)} \\ &= \frac{\binom{n_1}{x}\binom{n_2}{z-x}e^{x\theta}}{\sum_u \binom{n_1}{u}\binom{n_2}{z-u}e^{u\theta}},\end{aligned}$$

this distribution is so called the non-central hypergeometric distribution, where θ is the log odds ratio[12] and it doesn't depend on δ (we've conditioned on $X + Y$ and eliminated the nuisance parameter). This distribution can then be used to calculate exact hypothesis tests for $H_0 : \theta = \theta_0$. Inverting said exact test yields the exact confidence interval for the odds ratio. Note that this distribution simplifies to the hypergeometric distribution for $\theta = 0$ and the test is none other than Fisher's exact test. The R-routine `Fisher.test` performs this calculation.

2.4.7. *Matched 2x2 Tables.* We're now interested in treating matched two-by-two tables, which are similar in nature to paired T-tests but with binary data, the subject of dependence and the relationship to the CMH test.

We're interested in studying the approval and disapproval rates of a politician on two occasions. The dataset is given in 13.

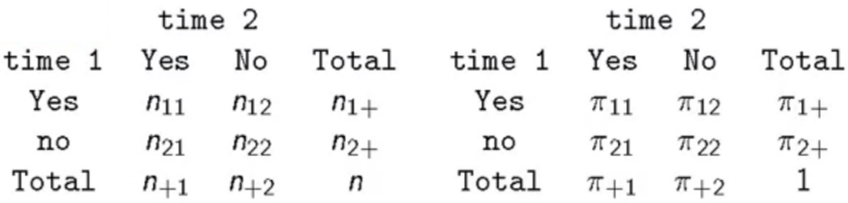
[12] Note this test is similar to Fisher's exact test but the difference lies in that we haven't assumed the null hypothesis to be true and depends on θ.



First survey	Second Survey Approve	Disapprove	Total
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

Controls	Cases Exposed	Unexposed	Total
Exposed	27	29	56
Unexposed	3	4	7
Total	30	33	63

FIGURE 13. Matched 2x2 tables Example from Agresti's book



time 1	time 2 Yes	No	Total
Yes	n_{11}	n_{12}	n_{1+}
no	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

time 1	time 2 Yes	No	Total
Yes	π_{11}	π_{12}	π_{1+}
no	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	1

FIGURE 14. Notation convention for matched 2x2 tables.

In this case, 749 had the same positive opinion in both surveys and 150 approved the first time and disapproved the second time. This example can be though as a retrospective case-control test.

2.4.8. *Dependence and Marginal Homogeneity.* Matched binary data can arise from

- Measuring a response at two occasions,
- matching on case status in a retrospective study,
- or by matching on exposure status in a prospective or cross-sectional study.

Matching in general induces dependence which need to be accounted for in the analysis. The pairs on binary observations are dependent, so our existing methods are no longer applicable. We'll discuss the process of making conclusions about the marginal probabilities and odds. We'll assume independence across pairs and dependence within pairs.

Let's consider a general example for matched tables and notations. Consider figure 14 and the standard contigency table notation given in table 1.

We assume that $(n_{11}, n_{12}, n_{21}, n_{22})$ are multinomial with n trials and success probabilities $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ respectively. This is, every time 1-time 2 collection pair of measurements, is going to be a one or a zero in exactly in one of these four locations. So the person will have either said yes at both occasions, a yes and no, a no and a yes or no at both occasions. The probability of being a one in the ijth-cell is π_{ij}. Note that π_{1+} and π_{+1} are the marginal probabilities of a yes response at the two occasions, this is $\pi_{1+} = \mathbb{P}(\text{Yes}|\text{Time 1})$ (regardless of Time 2's results) and $\pi_{+1} = \mathbb{P}(\text{Yes}|\text{Time 2})$ (regardless of Test 1's results). Similarly for the second set of π_2.

We define marginal homogeneity is the hypothesis $H_0 : \pi_{1+} = \pi_{+1}$ which is equivalent to the (transpose) symmetry of the off-diagonal elements $H_0 : \pi_{12} = \pi_{21}$[13]. The obvious estimate for $\pi_{12} - \pi_{21}$ is

$$\frac{n_{12}}{n} - \frac{n_{21}}{n},$$

[13] This statement is only true for 2-by-2 tables and doesn't stand for more general cases.

which quantifies how far away from symmetry the data is ie. how far away from marginal homogeneity the data is. Note that under the null hypothesis, a consistent estimate of the variance is $\frac{n_{12}+n_{21}}{n^2}$. Therefore the z-statistic is the previous estimate divided by the standard error ie.

$$\frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}} \underset{\text{asymp.}}{\sim} \chi^2_1.$$

This is the so-called **McNemar's test statistic**, and we reject the null hypothesis if the test statistic is large. Notice only discordant cells enter into the test, both n_{12} and n_{21} carry the relevant information about whether or nor π_{1+} and π_{+1} differ. Notice n_{11} and n_{22} contribute information to estimating the magnitude of this difference.

Consider the approval rating example. In that case, we have 86 and 150 as the off-diagonal cells, then the test statistic is $\frac{(86-150)^2}{86+150} = 17.36$, with a P-value of 3×10^{-5}. Hence, we reject the null hypothesis and conclude there is evidence to suggest a change in opinion between the two polls. This test can be performed via the following R routine

```
mcnemar.test(matrix(c(794, 86, 150, 570), 2),
             correct=FALSE)
```

which doesn't applies a continuity correction.

2.4.9. *Estimation of the Marginal Difference in Proportions.* We'd also like to construct a confidence interval for our test. Let $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$ be the estimated sample proportions and we're interested in estimating $d = \hat{\pi}_{1+} - \hat{\pi}_{+1} = \frac{n_{12}-n_{21}}{n}$, the difference in marginal proportions. The variance of d, in general (needing not the null hypothesis to be true) is

$$\sigma_d^2 = \frac{1}{n} \times \Big[\pi_{1+}(1-\pi_{1+}) + \pi_{+1}(1-\pi_{+1}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})\Big],$$

where the last term accounts for the correlation in the dataset, where $\frac{d-(\pi_{1+}-\pi_{+1})}{\hat{\sigma}_d} \underset{\text{asymp.}}{\sim} \mathcal{N}(0,1)$, from where it's trivial to construct the confidence interval.

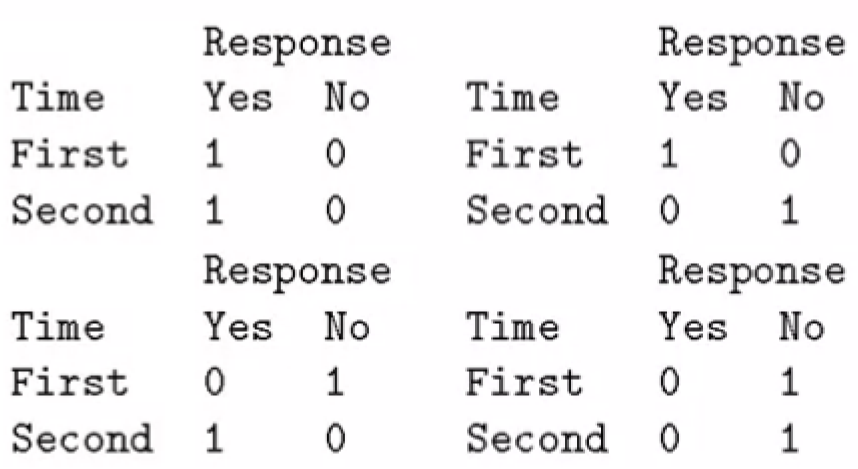
We could also compare σ_d^2 with what we would use if the proportions were independent, this is: instead of asking the same test subjects on two occasion whether or not they approve, what if we tested a different set of people each time? it would be logical to think that people who approved on the first occasion, would be more likely to approve on the second occasion, and the same logic applies to disapproval rates; the measurements tend to be concordant. This happens when $\pi_{11}\pi_{22} > \pi_{12}\pi_{21}$, tending to lie on the main diagonal, which implies a dramatic reduction in the variance. Failing to account for the fact that the same people were asked twice, would imply a much wider, less precise, confidence interval.

2.4.10. *Odds and Ends for Matched 2x2 Tables.* Now, let's go through our example:

- the difference in marginal proportions is $d = \frac{944}{1600} - \frac{880}{1600} = .59 - .55 = .04$,
- $\hat{\pi}_{11} = .50$, $\hat{\pi}_{12} = .09$, $\hat{\pi}_{21} = .05$, $\hat{\pi}_{22} = .36$,
- $\hat{\sigma}_d^2 = \frac{.59(1-.59)+.55(1-.55)-2\times(.50\times.36-.09\times.05}{1600}$, thus $\hat{\sigma}_d = .0095$.
- $95\%CI = .04 \pm 1.96 \times 0.095 = [.06, .02]$.

Note that ignoring dependence would yield $\hat{\sigma}_d = .0175$, giving a lower performing confidence interval.

There's a non-trivial relationship to the Cochran-Mantel-Haenszel test and the matched 2-by-2 tables. Consider the situation in which we take each pair and represented their time, first and second, and report their responses, either a yes or a no. This can be thought as an extremely stratified setting where every strata has the two measurements. Doing this leaves only four possible tables, pictured in 15. These tables will exactly reproduce the two-by-two table.



	Response			Response	
Time	Yes	No	Time	Yes	No
First	1	0	First	1	0
Second	1	0	Second	0	1

	Response			Response	
Time	Yes	No	Time	Yes	No
First	0	1	First	0	1
Second	1	0	Second	0	1

FIGURE 15. Relationship to the Cochran-Mantel-Haenszel test.

The McNemar's test is equivalent to the CMH test where subject is the stratifying variable and each 2×2 table is the observed zero-one table for that subject. This is useful only as a theoretical justification. Performing the CMH test on these heavily-stratified dataset, the tables, with the stratifying variable being the subjects results in McNemar's test.

McNemar's test has an exact version. Consider the off-diagonal cells, n_{12} and n_{21}. Under the null hypothesis $H_0 : \frac{\pi_{12}}{\pi_{12}+\pi_{21}} = .5$. Therefore, under H_0, $n_{21}|n_{12}+n_{21}$ is binomial with success probability .5 and $n_{12}+n_{21}$ trials. We can use this result to construct an exact P-value for matched pairs data. Under the null hypothesis, the two off-diagonal probabilities are identical. Landing on either of the off-diagonal cells is a fair coin-toss for every matched pair. We'd have evidence against H_0 if the said number of trials were not equal. This is related to the **non-parametric sine test**, under the null hypothesis, results should exchangeable whether they agree in terms of approving and disapproving or disapproving and approving.

Consider the approval rating data. Let $H_0 : \pi_{21} = \pi_{12}$ versus the alternative $H_a : \pi_{21} < \pi_{12} \equiv \pi_{+1} < \pi_{1+}$, where π_{+1} is the approval at time 2 disregarding time 1 whilst π_{1+} is the approval at time 1 disregarding time 2. We're testing if the approval at time 2 is lower that the approval at time 1. 86 people disapproved on the first survey and approved on the second survey (n_{21} cell). In particular, we're going to test if that is smaller than what would be expected by chance and find the probability of getting data as or more extreme in favour of the alternative case: $\mathbb{P}(X \leq 86|86+150) = .0$ where $X \sim \text{Bin}(p = .5, n = 236)$. We then reject the null hypothesis. For two sided tests, we can double the smaller of the two one-sided tests.

The marginal odds ratio is the odds of approval at time 1 relative to the odds of approval at time 2:

$$\frac{\pi_{1+}/\pi_{2+}}{\pi_{+1}/\pi_{+2}} = \frac{\pi_{1+}\pi_{+2}}{\pi_{+1}\pi_{2+}}.$$

which is the quotient of the odds of approval at time 1 divided by the odds of approval at time 2 in the denominator, its a marginal odds ratio since we're dealing with marginal probabilities. Note this is in essence a paired-group test since the same sample group was tested twice. The maximum likelihood estimate of the marginal log odds ratio is

$$\hat{\theta} = \log \frac{\pi_{1+}\pi_{+2}}{\pi_{+1}\pi_{2+}},$$

and the asymptotic variance of this estimator is

$$\frac{1}{\pi_{1+}\pi_{2+}} + \frac{1}{\pi_{+1}\pi_{+2}} - 2\frac{(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})/(\pi_{1+}\pi_{2+}\pi_{+1})\pi_{+2}}{n},$$

which can then be used to calculate a confidence interval for a two by two dataset.

In the approval rating example, the marginal OR compares the odds of approval at time 1 to that at time 2. Then $\hat{\theta} = \log(944 \times 720/(880 \times 656)) = .16$, with an estimated standard error of .039. Therefore the confidence interval for the log odds ratio is $.16 \pm 1.96 \times .039 = [.084, .236]$.

Let's imagine if we created a logit model for our approval rating data.

$$\text{logit}(\mathbb{P}(\text{Person } i \text{ says Yes at Time 1}) = \alpha + U_i$$
$$\text{logit}(\mathbb{P}(\text{Person } i \text{ says Yes at Time 2}) = \alpha + \gamma + U_i,$$

where each U_i contains person-specific random effects. A person with a large U_i is likely to answer Yes at both occasions, where γ is the **log odds ratio** comparing a response of Yes at Time 1 to a response of Yes at Time 2, which is a subject specific effect. If we subtract the log odds of a yes responde for two different people, the U_i terms wouldn't cancel. One way to eliminate U_i is to perform a conditional estimate, and to condition on the total number of Yes responses for each person: 1. if they answered yes or no on both ocassion then you know both responses, therefore only discordant pairs have any relevant information after conditioning. Therefore, the conditional ML estimate for γ and its standard error are

$$\log \frac{n_{21}}{n_{12}} \sqrt{\frac{1}{n_{21}} + \frac{1}{n_{21}}}.$$

The marginal ML has a marginal interpretation. The effect is averaged over all of the values of U_i. The conditional ML estimate has a subject specific interpretation. Marginal interpretations are in general more useful for policy type statements. Policy makers tend to be interested in how factors influence populations. Subject specific interpretations are more useful in clinical applications. Physicians are interested in how many factor influence the individuals.

2.4.11. *The Sign Test.* In this section we'll treat some non-parametric tests, which include the sign test, the signed rank test (useful for paired tests), Monte-Carlo simulations of said test, independent groups, the Mann-Whitney test and their relationship to permutation tests.

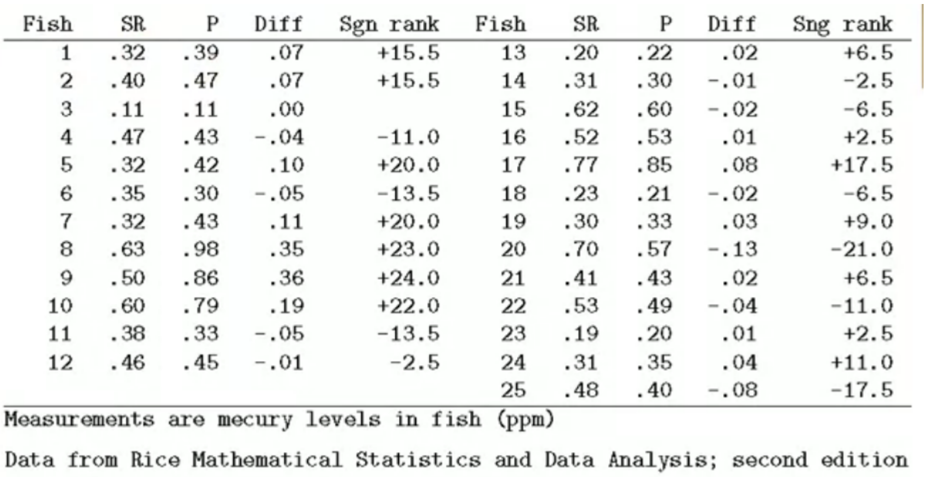
In general, distribution-free methods require fewer assumptions than parametric methods. In said types of test, the focus is on testing rather than estimation and are not sensitive to outlying observation. These are specially suited to analyse cruder data, like ranks, and discard some of the information in the data, which is a downside. Therefore they may be less powerful than parametric counterparts, even when the parametric assumptions are true. For large-enough samples, they are equally efficient to parametric counterparts.

Consider the following example, where 25 fish were tested at two locations for their mercury levels in parts per million.

In this dataset, since the same fish were tested at two different locations, each fish serves as its own control. Therefore with a non-parametric test we're interested in analysing the validity of the assumptions, such as normality.

Let $D_i = \text{difference}(P - SR)$, let θ be the population median of the difference D_i. The null hypothesis is $H_0 : \theta = 0$ versus $H_a : \theta \neq 0$, or one of the other two alternatives. By definition of the median, $\theta = 0$ if and only if $p = \mathbb{P}(D > 0) = .5$. Assuming iid data, let X be the number of times $D > 0$ which in turn implies $X \sim \text{Binom}(n, p)$. The sign tests whether $H_0 : p = .5$ using X, if X is excessively large or excessively small, this would dispute the null hypothesis, by performing an exact binomial test.

Consider the previous example. Let $\theta = \text{median difference}\, p - sr$, with the null hypothesis being $H_0 : \theta = 0$ versus the alternative hypothesis being $H_a : \theta \neq 0$. The number of instances where the difference is bigger than 0 is 15 out 25 trials. From the R-function, we can test if 15 out 25 trials is excessively larger than 12.5 `binom.test(15,25)` which gives the `p-value = .4244`, thus there's a 42%



Fish	SR	P	Diff	Sgn rank	Fish	SR	P	Diff	Sng rank
1	.32	.39	.07	+15.5	13	.20	.22	.02	+6.5
2	.40	.47	.07	+15.5	14	.31	.30	-.01	-2.5
3	.11	.11	.00		15	.62	.60	-.02	-6.5
4	.47	.43	-.04	-11.0	16	.52	.53	.01	+2.5
5	.32	.42	.10	+20.0	17	.77	.85	.08	+17.5
6	.35	.30	-.05	-13.5	18	.23	.21	-.02	-6.5
7	.32	.43	.11	+20.0	19	.30	.33	.03	+9.0
8	.63	.98	.35	+23.0	20	.70	.57	-.13	-21.0
9	.50	.86	.36	+24.0	21	.41	.43	.02	+6.5
10	.60	.79	.19	+22.0	22	.53	.49	-.04	-11.0
11	.38	.33	-.05	-13.5	23	.19	.20	.01	+2.5
12	.46	.45	-.01	-2.5	24	.31	.35	.04	+11.0
					25	.48	.40	-.08	-17.5

Measurements are mecury levels in fish (ppm)

Data from Rice Mathematical Statistics and Data Analysis; second edition

Figure 16. Dataset for the Sign test.

chance of happening under the null hypothesis for a two sided test. Or we could have used large sample tests for a binomial proportion:

```
prop.test(15, 25, p=.5)
X-squared = .64, df = 1, p-value = 0.4237
```

In essence this test analyses if levels of one data group are higher than the levels of the other group, by counting the number of pairs from the matched pairs where it's higher and by quantifying if this number is excessively large relative to a coin flip for each pair.

The only assumption made on the data is that the data-points are iid. There are some problems with these kind of problems. The magnitude of the difference itself is discarded, perhaps too much information being lost in the process. We could easily test $H_0 : \theta = \theta_0$, for any value of θ_0, by calculating the number of times $D > \theta_0$ and performing a binomial test. We can invert these test to get a distribution free confidence interval for the median.

2.4.12. *The Sign Rank Test.* The main issue with the sign test is that the magnitude of differences is discarded, the test potentially not being as powerful as other techniques.

Wilcoxon treated the problem of discarding the magnitude of the differences by creating a test statistic which also takes into account the signed ranks of the differences, saving some of the information regarding the magnitude of the differences. The null hypothesis $H_0 : \theta = 0$ is thus tested against one of the three alternatives. Appropriately normalised, the test statistic follows a normal distribution. Also note the exact small sample distribution of the signed rank statistic is known if there are no ties. With this test we can quantise how "normal" our distribution is. The signed rank procedure can be implemented as follows:

- Take the paired differences,
- take the absolute values of the differences,
- rank these absolute values and throwing away the zeros (throwing away the ties).
- Then, multiply the ranks by the sign of the difference,
- and calculate the rank sum W_+ of the positive ranks

Therefore, remembering θ is the median,

- If $\theta > 0$, then W_+ should be large.
- If $\theta < 0$ then W_+ should be small.

Properly normalized $W_+ \sim \mathcal{N}(\mu, \sigma^2)$ for some $\mu, \sigma \in \mathbb{R}$. For small sample sizes W_+ has an exact distribution under the null hypothesis, being able to extract critical values from standardises tables.

This test can be performed with Monte Carlo simulations by simulating n observation from any distribution symmetric around zero, this is with $\theta = 0$ as its median, getting the small sample distribution out of this procedure[14]. Then, we rank the absolute value of the data, retain the signs, and calculate the signed rank statistic. Applying this procedure a great many number of times, the proportion of time that the observed statistic is larger or smaller is a Monte Carlo approximation to the P-value, depending on the hypothesis. Consider the ranks $1, \cdots, n$. Randomly assign the signs as binary with probability .5 of being positive and .5 of being negative. Then, by calculating the signed rank statistic and applying this produce for many simulations, the proportion of times that the observed test statistic is larger or smaller is a Monte Carlo approximation to the P-value.

We're interested in studying the large sample distribution of W_+. Under H_0 and if there are no ties,

$$\mathbb{E}(W_+) = \frac{n(n+1)}{4},$$
$$\mathbb{V}\mathrm{ar}(W_+) = \frac{n(n+1)(2n+1)}{24},$$
$$TS = \frac{W_+ - \mathbb{E}(W_+)}{SD(W_+)} \rightarrow \mathcal{N}(0,1).$$

There is a correction term necessary to account for ties. Without ties, it's possible however to perform an exact small sample test.

A Wilcoxon test can be performed with the following R routine

```
diff <- c(.07, .07, .00, -.04,...)
wilcox.test(diff,exact=FALSE)

In [1]: Wilcoxon signed rank test with continuity correction

In [2]: data:  diff
In [3]: V = 5, p-value = 0.4142
In [4]: alternative hypothesis: true location is not equal to 0
```

In this case, since the P-value is large, we cannot rule out the null hypothesis.

2.4.13. *The Rank Sum Test.* The sign ranked test is the non-parametric equivalent to the paired test. Now we'll treat the non-parametric equivalent to the unpaired test. Consider the following example, in which we're comparing two measuring techniques A and B, where the units are given in degº C per gram, where we're interested in testing whether the two techniques produce equivalent results.

Method A		Method B
79.98	80.05	80.02
80.04	80.03	79.94
80.02	80.02	79.98
80.04	80.00	79.97
80.03	80.02	79.97
80.03		80.03
80.04		79.95
79.97		79.97

[14] Note that under the null hypothesis, the signs are equally likely to be distributed anywhere among the ranks. Then we'll take the ranks of numbers between 1 and n and randomly allocate the signs (by essentially flipping a coin) which gives the null distribution of the signed rank statistic

The basis of thought for this index is to take the AB-labels and shuffle them on every measurement.

The Mann-Whitney test/Wilcoxon rank sum test tests whether or not the two treatments have the same location (ie. tests whether they are centred at the same place) by assuming iid errors, not necessarily normal. The null hypothesis can also be written more generally as a stochastic shift for two arbitrary distributions (the distribution for method B is uniformly shifted relative to distribution A). This test uses the sum of the ranks obtained by discarding the treatment labels AB. In order words, we have to perform the following steps

- Discard the treatment labels,
- Rank the observations,
- Calculate the sum of the ranks in the first treatment and
  - either calculate the asymptotic normal distribution of this statistic,
  - or compare with the exact distribution under the null hypothesis.

Continuing our previous example we have that the rank sum table is

Method A	Method B
7.5 21.0	11.5
19.0 15.5	1.0
11.5 11.5	7.0
19.0 9.0	4.5
15.5 11.5	4.5
15.5	15.5
19.0	2.0
4.5	4.5
180	51

Note that the sum has to add up to $21 \times 22/2 = 231$.

Let W be the sum of the ranks for the first treatment A. Note that as the sample size grows, by definition, W also grows as well. Let n_A and n_B be the sample sizes, which are known, then

$$\begin{aligned}\mathbb{E}(W) &= \frac{n_A(n_A + n_b + 1)}{2},\\ \mathbb{V}\text{ar}(W) &= \frac{n_A n_B(n_A + n_B + 1)}{12},\\ TS =& \frac{W - \mathbb{E}(W)}{SD(W)} \to \mathcal{N}(0,1), .\end{aligned}$$

where all the calculations are performed under the null hypothesis. Note that the exact distribution for W can be calculated.

In our previous example, if we analyse by method B, $W = 51$, $\mathbb{E}(W) = (8 + 13 + 1)/2 = 88$, where the standard error is $SD(W) = \sqrt{8 \times 13(8 + 13 + 1)/12} = 13.8$ and the test statistic is $TS = (51 - 88)/13.8 = -2.68$. It turns out the two-sided P-value is .007 thus we reject the null hypothesis. This calculation can be performed in R with the function `wilcox.test`.

Now, let's note that under the null hypothesis H_0 the two groups are exchangeable. Therefore, any allocation of the ranks between the two groups is equally likely. Thus it's possible to perform a Monte Carlo simulation of the the experiment by performing the following procedure

- Take the ranks $1, \cdots, N_A + N_B$ and permute them,
- take the first N_A ranks and allocate them to Group A and allocate the remainder to group B.
- Calculate the test statistic,

- and repeat the process over and over, the proportion of times the test statistic is larger or smaller (depending on the alternative) than the observed value is an exact P-value.

We're now interested in considering permutation test, which are similar in nature to the rank-sum tests, though the use the actual data rather than the ranks. That is, consider the null hypothesis that the distribution of the observations from each group are the same. Then, the labels are irrelevant, we can discard them and permute the combined data by splitting the permute data into two groups, group A and group B, with n_A and n_B observation respectively. By evaluating the probability of getting a statistic as large or larger than the one observed gives the P-value. An example statistic would be the difference between the averages of the two groups or one could perform a T-test. Note that Fisher's exact test (which works on binary data), the rank sum test (which works on ranking the observations), and the permutation tests all share the same basic principle: that the labels are exchangeable and our null distribution is obtained by permuting those labels across the values. In Fisher's exact test and the permutation test we permute them across the actual observed values and then in the rank sum test, we're converting the data to ranks first and then permuting across the ranks. This is an easy way to produce a null distribution for a test of equal distributions and is similar in nature to the bootstrap. This produces an exact test, though less robust but more powerful than the rank sum test and is widely used in genomic applications.

2.4.14. *Poisson Distribution.* The Poisson distribution is a useful model for analysing event/time datasets, for the modelling of radioactive decay, the modelling of survival data (survival analysis being modelling the time until some event, typically death, occurs, for example in the study of diseases), any sort of model for unbounded count data, for contingency tables and as an approximation of binomial random variables when n is large and p is small.

2.4.15. *Poisson Likelihood.*

2.4.16. *Poisson P-value Calculation.*