Thinking About Data

Ryan G

March 20, 2020

Contents

1 TODO Overheads

1.0.1 TODO Install Emacs Application Framework

This is going to be necessary to deal with not just equations but links, tables and other quirks

Install it from here

The reason for this is that generating latex preview fragments is just far too slow to be useful in any meaningful fashion.

1.0.2 TODO Install a live preview for equations in org-mode

Here is one example but there was a better one I was using

2 Deriving the Normal Distribution

2.1 Power Series

Series

A function f :

$$f\left(j\right) = \sum_{i=0}^{\infty} \left[C_n \left(z-a\right)^n\right], \ \exists z \in \mathbb{C}$$

$$f\left(z\right) = \sum_{i=0}^{\infty} \left[C_n \left(z-a\right)^n\right], \ \exists z \in \mathbb{C}$$

Is a Power Series a and will either:

- Converge only for $x = a$,
- converge $\forall x$
- converge in the circle $|z-a| < R$

2.1.1 Example

Take some function equal to the following power series:

$f\left(x\right) = \sum_{n=0}^{\infty} \left[n! \cdot x^n\right]$

Because the terms inside the power series has a factorial the only test that will work is the limit ratio test so we use that to evaluate convergence. [1]

let $a_n = n! \cdot x^n$:

$$\begin{aligned}\frac{\lim_{n\to\infty}|a_{n+1}|}{\lim_{n\to\infty}|a_n|} &= \lim_{n\to\infty}\left|\frac{(n+1)! \cdot x^n \cdot x}{n! \cdot x^n}\right| \\ &= (n+1)\cdot|x| \\ &= 0 \iff x = 0\end{aligned}$$

$\therefore$ The power series converges if and only $x = 0$.

2.1.2 Representing a function as a Power Series

Ordinary functions can be represented as power series, this can be useful to deal with integrals that don't have an elementary anti-derivative.

[1] Refer to Solving Series Strategy

1. Geometric Series First take the Series:

$$\begin{aligned} S_n &= \sum_{n=0}^{n} r^k \\ &= 1 + r + r^2 + r^3 \ldots + r^{n-1} + r^n \\ \implies r \cdot S_n &= r + r^2 + r^3 + r^4 \ldots r^n + r^{n+1} \\ \implies S_n - r \cdot C_n &= 1 + r^{n+1} \\ \implies S_n &= \frac{1 + r^{n+1}}{1 - r} \end{aligned}$$

So now consider the geometric series:

$$\begin{aligned} \sum_{k=0}^{\infty} \left[x^k \right] &= \lim_{n \to \infty} \left[\sum_{k=0}^{n} x^k \right] \\ &= \lim_{n \to \infty} \left[\frac{1 + x^{n+1}}{1 - x} \right] \\ &= \frac{1 + \lim_{n \to \infty} \left[x^{n+1} \right]}{1 - x} \\ &= \frac{1 + 0}{1 - x} \\ &= \frac{1}{1 - x} \end{aligned}$$

2. Using The Geometric Series to Create a Power Series Take for example the function:

$$g(x) = \frac{1}{1 + x^2}$$

This could be represented as a power series by observing that:

$$\frac{1}{1 - \#_1} = \sum_{n=0}^{\infty} [\#_1^n]$$

And then simply putting in the value of $\#_1 = \left(-x^2\right)$:

$$\frac{1}{1 - (-x^2)} = \sum_{n=0}^{\infty} \left[\left(-x^2\right)^n \right]$$

2.1.3 Calculus Rules and Series

The laws of differentiation allow the following relationships:

1. Differentiation

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\sum_{n=1}^{\infty} c_n \left(z-a\right)^n\right) = \sum_{n=1}^{\infty}\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(c_n \left(z-a\right)^n\right)\right]$$

2. Integration

$$\int\left(\sum_{n=1}^{\infty} c_n \left(z-a\right)^n\right)\mathrm{d}x = \sum_{n=1}^{\infty}\left[c_n \left(z-a\right)^n\right]$$

2.1.4 Taylor Series

This is the important one, the idea being that you can use this to easily represent any function as an infinite series:

Consider the pattern formed by taking derivatives of $f\left(z\right) = \sum_{n=1}^{\infty} c_n \left(z-a\right)^n$:

$$\begin{aligned}
f\left(z\right) &= c_0 + c_1\left(x-a\right) + c_2\left(x-a\right)^2 + c_3\left(x-a\right)^3 + \ldots \\
&\implies f\left(a\right) = c_0 \\
f'\left(z\right) &= c_1 + 2c_2\left(z-a\right) + 3c_3\left(z-a\right)^2 + 4c_4\left(z-a\right)^3 \\
&\implies f'\left(a\right) = c_1 \\
f''\left(z\right) &= 2c_2 + 3\times 2\times c_3\left(z-a\right) + 4\times 3c_4\left(z-a\right)^2 + \ldots \\
&\implies f''\left(a\right) = 2\cdot c_2 \\
f'''\left(z\right) &= 3\times 2\times 1\cdot c_3 + 4\times 3\times 2c_4\left(z-a\right) + \ldots \\
&\implies f'''\left(a\right) = 3!c_3
\end{aligned}$$

Following this pattern forward:

$$\begin{aligned}
&f^{(n)}\left(a\right) = n!\cdot c_n \\
&\implies c_n = \frac{f^{(n)}\left(a\right)}{n!}
\end{aligned}$$

Hence, if there exists a power series to represent the function f, then it must be:

$$f\left(z\right) = \sum_{n=0}^{\infty}\left[\frac{f^{(n)}\left(a\right)}{n!}\left(x-a\right)^n\right]$$

If the power series is centred around 0, it is then called a *Mclaurin Series.*

1. Power Series Expansion of e

$$\begin{aligned} f(z) = e^z &= \sum_{n=0}^{\infty}\left[\frac{f^{(n)}(0)}{n!}\cdot x^n\right] \\ &= \sum_{n=0}^{\infty}\left[\frac{e^0}{n!}x^n\right] \\ &= \sum_{n=0}^{\infty}\left[\frac{x^n}{n!}\right] \end{aligned}$$

2.2 Modelling Normal Distribution

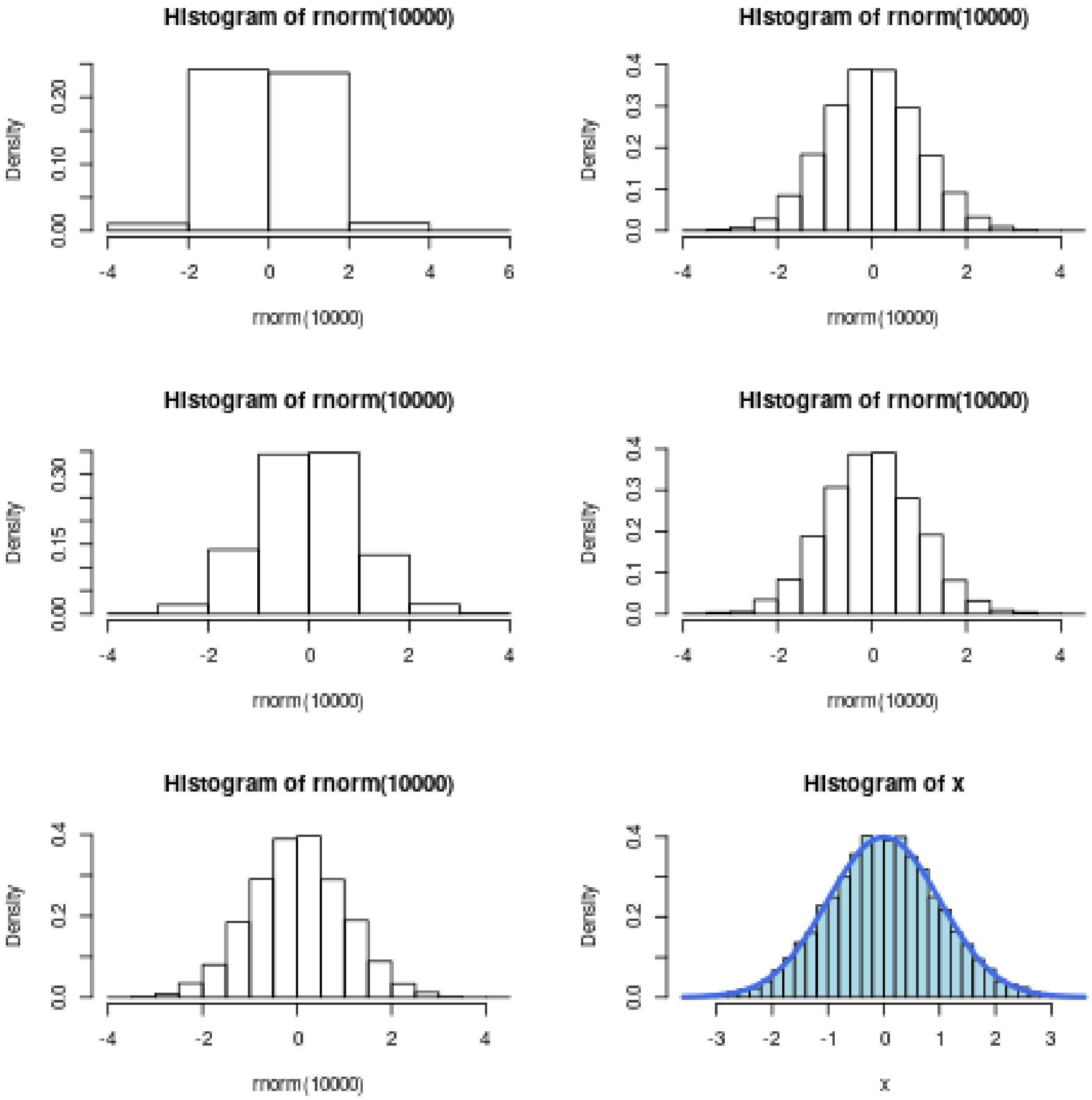
The Normal Distribution is a probability density function that is essentially modelled after observation.[2]

2.2.1 what is the $y$-axis in a Density curve?GGPLOT2:ATTACH

Consider a histogram of some continuous normally distributed data:

```
layout(mat = matrix(1:6, nrow = 3))

x <- rnorm(10000, mean = 0, sd = 1)
sd(x)
hist(rnorm(10000), breaks = 5, freq = FALSE)
## curve(dnorm(x, 0, 1), add = TRUE, lwd = 3, col = "royalblue")

hist(rnorm(10000), breaks = 10, freq = FALSE)
## curve(dnorm(x, 0, 1), add = TRUE, lwd = 3, col = "royalblue")

hist(rnorm(10000), breaks = 15, freq = FALSE)
## curve(dnorm(x, 0, 1), add = TRUE, lwd = 3, col = "royalblue")

hist(rnorm(10000), breaks = 20, freq = FALSE)
## curve(dnorm(x, 0, 1), add = TRUE, lwd = 3, col = "royalblue")

hist(rnorm(10000), breaks = 25, freq = FALSE)
## curve(dnorm(x, 0, 1), add = TRUE, lwd = 3, col = "royalblue")

hist(x, breaks = 30, freq = FALSE, col = "lightblue")
curve(dnorm(x, 0, 1), add = TRUE, lwd = 3, col = "royalblue")
```

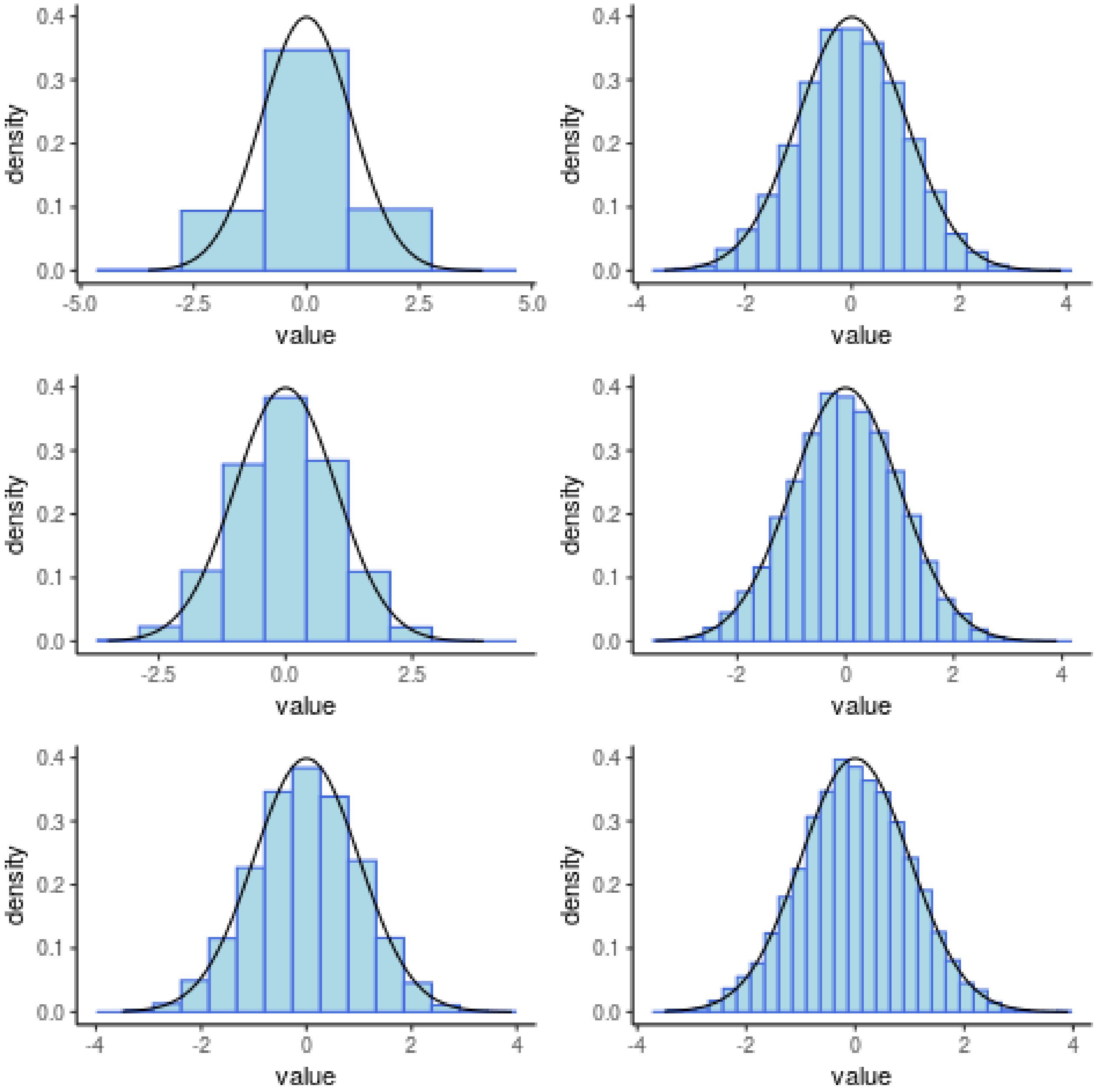
[2] The Normal Distribution



(Or in ggplot2):

```
library(tidyverse)
```

```r
library(gridExtra)
x <- rnorm(10000)
x <- tibble::enframe(x)
head(x)
PlotList <- list()
for (i in seq(from = 5, to = 30, by = 5)) {
  PlotList[[i/5]] <- ggplot(data = x, mapping = aes(x = value)) +
    geom_histogram(aes(y = ..density..), col = "royalblue", fill = "lightblue", bins =
    stat_function(fun = dnorm, args = list(mean = 0, sd = 1))+
    theme_classic()
}

                                        # arrangeGrob(grobs = PlotList, layout_matrix
grid.arrange(grobs = PlotList, layout_matrix = matrix(1:6, nrow = 3))
```



Observe that the outline of the frequencies can be made arbitrarily close to a curve given that the bin-width is made sufficiently small. This curve, known as the probability density function, represents the frequency of observation around that value, or more accurately the area beneath the curve around that point on the $x$-axis will be the probability of observing values within that corresponding interval.

Strictly speaking the curve is the rate of change of the probability at that point as well.

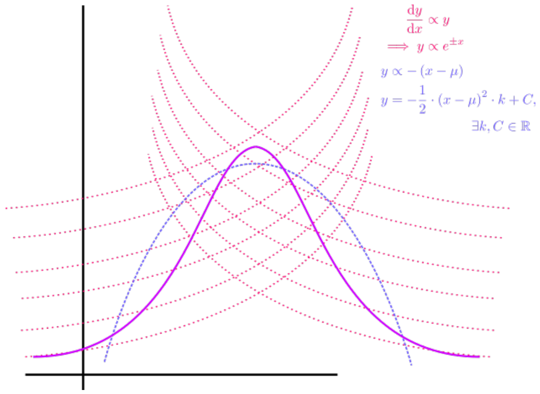
2.2.2 Defining the Normal Distribution

Data are said to be normally distributed if, the plot of the frequency density curve is such that:

- The rate of change is proportional to:
  - The distance of the score from the mean
    - $\frac{\mathrm{d}}{\mathrm{d}x}(f) \propto -(x-\mu)$
  - The frequencies themselves.
    - $\frac{\mathrm{d}}{\mathrm{d}x} \propto f$

If the Normal Distribution was only proportional to the distance from the mean (i.e. $(x \propto -(x-\mu))$ the model would be a parabola that dips below zero, as shown in 2.2.3, so it is necessary to provide the restriction that the rate of change is also proportional to the frequency (i.e. $y \propto y$).

let f be the frequency of observation around $x$, following these rules the plot would come to look something like:



2.2.3 Modelling only distance from the mean

If we presumed the frequency (which we will call f on the $y$-axis) was proportional only to the distance from the mean the model would be a parabola:

$$
\begin{aligned}
\frac{\mathrm{d}f}{\mathrm{d}x} &\propto -\left(x-\mu\right) \\
\frac{\mathrm{d}f}{\mathrm{d}x} &= -k\left(x-\mu\right), \quad \exists k \in \mathbb{R} \\
\int \frac{\mathrm{d}f}{\mathrm{d}x}\mathrm{d}x &= -\int \left(x-\mu\right)\mathrm{d}x
\end{aligned}
$$

Using integration by substitution:

$$
\begin{aligned}
\text{let:} \quad v &= x-\mu \\
\implies \frac{\mathrm{d}v}{\mathrm{d}x} &= 1 \\
\implies \mathrm{d}v &= \mathrm{d}x
\end{aligned}
$$

and hence

$$
\begin{aligned}
\int \frac{\mathrm{d}f}{\mathrm{d}x}\mathrm{d}x &= -\int \left(x-\mu\right)\mathrm{d}x \\
\implies \int \mathrm{d}p &= -\int v\mathrm{d}v \\
p &= -\frac{1}{2}v^2 \cdot k + C \\
p &= -\frac{1}{2}\left(x-\mu\right)^2 \cdot k + C
\end{aligned}
$$

Clearly the problem with this model is that it allows for probabilities less than zero, hence the model needs to be refined to:

- incorporate a slower rate of change for smaller values of f (approaching 0)
- incorporate a faster rate of change for larger values of f
  - offset by the the condition that $\frac{\mathrm{d}f}{\mathrm{d}x} \propto -\left(x-\mu\right)$

2.2.4 Incorporating Proportional to Frequency

In order to make the curve bevel out for smaller values of f it is sufficient to implement the condition that $\frac{\mathrm{d}f}{\mathrm{d}x} \propto f$:

$$
\begin{aligned}
\frac{\mathrm{d}f}{\mathrm{d}x} &\propto f \\
\int \frac{1}{f}\cdot\frac{\mathrm{d}f}{\mathrm{d}x}\mathrm{d}x &= k\cdot\int \mathrm{d}x \\
\ln|f| &= k\cdot x \\
f &= C\cdot e^{\pm x} \\
f &\propto e^{\pm x}
\end{aligned}
$$

2.2.5 Putting both Conditions together

So in order to model the bell-curve we need:

$$
\begin{aligned}
f \propto f \wedge f &\propto -(x-\mu) \\
\implies \frac{\mathrm{d}f}{\mathrm{d}x} &\propto -f(x-\mu) \\
\int \frac{1}{f}\mathrm{d}f &= -k\cdot\int (x-\mu)\,\mathrm{d}x \\
\ln|f| &= -k\int (x-\mu)\,\mathrm{d}x
\end{aligned}
$$

because $f > 0$ by definition, the absolute value operators may be dispensed with:

$$
\begin{aligned}
\ln(f) &= -k\cdot\frac{1}{2}(x-\mu)^2 + C \\
f &\propto e^{\frac{(x-\mu)^2}{2}}
\end{aligned}
$$

Now that the function has been solved it is necessary to apply the IC's in order to further simplify it.

1. IC, Probability Adds to 1 The area bound by the curve must be 1 because it represents probability, hence:

$$
\begin{aligned}
1 &= \int_{-\infty}^{\infty} f\mathrm{d}f \\
1 &= -C\int_{-\infty}^{\infty} e^{\frac{k}{2}(x-\mu)^2}\mathrm{d}f
\end{aligned}
$$

   Using integration by substitution:

$$\begin{aligned}\text{let:}\quad u^2 &= \frac{k}{2}\left(x-\mu\right)^2\\ u &= \sqrt{\frac{k}{2}}\left(x-\mu\right)\\ \frac{\mathrm{d}u}{\mathrm{d}x} &= \sqrt{\frac{k}{2}}\end{aligned}$$

hence:

$$\begin{aligned}1 &= -C\int_{-\infty}^{\infty} e^{\frac{k}{2}(x-\mu)^2}\\ 1 &= \sqrt{\frac{2}{k}}\cdot C\int_{-\infty}^{\infty} e^{-u^2}\mathrm{d}u\\ 1^2 &= \left(\sqrt{\frac{2}{k}}\cdot C\int_{-\infty}^{\infty} e^{-u^2}\mathrm{d}u\right)^2\\ 1^2 &= \left(\sqrt{\frac{2}{k}}\cdot C\int_{-\infty}^{\infty} e^{-u^2}\mathrm{d}u\right)\times\left(\sqrt{\frac{2}{k}}\cdot C\int_{-\infty}^{\infty} e^{-u^2}\mathrm{d}u\right)\end{aligned}$$

Because this is a definite integral u is merely a dummy variable and instead we can make the substitution of x and y for clarity sake.

$$1^2 = \left(\sqrt{\frac{2}{k}}\cdot C\int_{-\infty}^{\infty} e^{-x^2}\mathrm{d}x\right)\times\left(\sqrt{\frac{2}{k}}\cdot C\int_{-\infty}^{\infty} e^{-y^2}\mathrm{d}y\right)$$

Now presume that the definite integral is equal to some real constant $\beta \in \mathbb{R}$:

$$\begin{aligned}1 &= \frac{2}{k}\cdot C^2\int_{-\infty}^{\infty} e^{-y^2}\mathrm{d}y\times\beta\\ &= \frac{2}{k}\cdot C^2\int_{-\infty}^{\infty}\beta\cdot e^{-y^2}\mathrm{d}y\\ &= \frac{2}{k}\cdot C^2\cdot\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} e^{-x^2}\mathrm{d}x\right)e^{-y^2}\mathrm{d}y\\ &= \frac{2}{k}\cdot C^2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\left(x^2+y^2\right)}\mathrm{d}x\mathrm{d}y\end{aligned}$$

This integral will be easier to evaluate in polar co-ordinates, a double integral may be evaluated in polar co-ordinates using the following relationship: [3]

$$\iint_D f(x,y)\,dA = \int_\alpha^\beta \int_{h_1(\phi)}^{h_2(\phi)} f(r\cdot\cos(\phi), r\cdot\sin(\phi))\,\mathrm{d}r\mathrm{d}\phi$$

hence this simplifies to:

$$\begin{aligned}1 &= \frac{2}{k}c^2 \int_0^{2\pi}\int_0^r r\cdot e^{(r\cdot\cos\theta)^2+(r\cdot\sin\theta)^2}\mathrm{d}r\mathrm{d}\theta \\ 1 &= \frac{2}{k}c^2 \int_0^{2\pi}\int_0^r r\cdot e^{r^2}\mathrm{d}r\mathrm{d}\theta\end{aligned}$$

Because the integrand is of the form $f'(x) \times g(f(x))$ we may use integration by substitution:

$$\begin{aligned}\text{let:}\quad u &= -r^2 \\ \frac{\mathrm{d}u}{\mathrm{d}r} &= -2r \\ \mathrm{d}r &= -\frac{1}{2r}\mathrm{d}u\end{aligned}$$

and hence:

$$\begin{aligned}1 &= \frac{2}{k}c^2 \int_0^{2\pi}\int_0^r r\cdot e^{r^2}\mathrm{d}r\mathrm{d}\theta \\ \implies 1 &= -\frac{2}{k}c^2 \int_0^{2\pi}\int_0^\infty r\cdot e^{r^2}\mathrm{d}r\mathrm{d}\theta\end{aligned}$$

[3] Calculus III - Double Integrals in Polar Coordinates

$$
\begin{aligned}
1 &= \frac{2}{k}c^2 \int_0^{2\pi} \int_0^{r} r \cdot e^{r^2} \mathrm{d}r \mathrm{d}\theta \\
\implies 1 &= -\frac{2}{k}c^2 \int_0^{2\pi} \int_0^{\infty} -\frac{1}{2} e^{-u} \mathrm{d}u \mathrm{d}\theta \\
&= \frac{1}{k}c^2 \int_0^{2\pi} \int_0^{\infty} e^{-u} \mathrm{d}u \mathrm{d}\theta \\
&= \frac{1}{k}c^2 \int_0^{2\pi} \left[-e^{-u}\right]_0^{\infty} \mathrm{d}\theta \\
1 &= \frac{1}{k}c^2 2\pi \\
\implies C^2 &= \frac{k}{2\pi}
\end{aligned}
$$

So from before:

$$
\begin{aligned}
f &= -C \cdot e^{k \cdot \frac{(x-\mu)^2}{2}} \\
&= -\sqrt{\frac{k}{2\pi}} \cdot e^{k \cdot \frac{(x-\mu)^2}{2}}
\end{aligned}
$$

so now we simply need to apply the next initial condition.

2. IC, Mean Value and Standard Deviation

   (a) Definitions The definition of the expected value, where $f(x)$ is a probability function is: [4]

   $$\mu = E(x) = \int_a^b x \cdot f\left(x\right) \mathrm{d}x$$

   That is, roughly, the sum of the expected proportion of occurence. The definition of the variance is:

   $$V\left(x\right) = \int_a^b \left(x - \mu\right)^2 f\left(x\right) \mathrm{d}x$$

   which can be roughly interpreted as the sum of the proportion of squared distance units from the mean. The standard deviation is $\sigma = \sqrt{V(x)}$.

[4] Expected Value and Variance

(b) Expected Value of the Normal Distribution The expected value of the normal distribution is μ, this can be shown rigorously:

$$\begin{aligned} \text{let:} \quad & v = x - \mu \\ \implies & \mathrm{d}v = \mathrm{d}x \end{aligned}$$

Observe that the limits of integration will also remain as $\pm\infty$ following the substitution:

$$\begin{aligned} E\left(v\right) &= \int_{-\infty}^{\infty} v \times f\left(v\right) \mathrm{d}v \\ &= k \cdot \int_{-\infty}^{\infty} v \cdot e^{v^2} \mathrm{d}v \\ &= \frac{1}{2} \left[e^{x^2} \right]_{\infty}^{\infty} \\ &= \frac{1}{2} \lim_{b \to \infty} \left[\left[e^{x^2} \right]_{-b}^{b} \right] \\ &= \frac{1}{2} \lim_{b \to \infty} \left[e^{b^2} - e^{(-b)^2} \right] \\ &= \lim_{b \to \infty} \left[0\right] \times \frac{1}{2} \\ &= \frac{1}{2} \times 0 \\ &= 0 \end{aligned}$$

Hence the Expected value of the standard normal distribution is $0 = x - \mu$ and so $E(x) = \mu$.

(c) Variance of the Normal Distribution Now that the expected value has been confirmed, consider the variance of the distribution:

$$\sigma^2 = \int_{-\infty}^{\infty} \left(x - \mu\right)^2 \times f\left(x\right) \mathrm{d}x$$

Now observe that $(x - \mu)$ appears as an exponential and as a factor if this is redefined as $w = x - \mu \implies \mathrm{d}x = \mathrm{d}w$ we have:

$$\sigma^2 = \sqrt{\frac{k}{2}} \int_{-\infty}^{\infty} w^2 e^{-\frac{k}{2} w^2} \mathrm{d}w$$

Now the integrand is of the form $f\left(x\right) \times g\left(x\right)$ meaning that the only strategy to potentially deal with it is integration by parts:

$$\int u \mathrm{d}v = u \cdot v - \int v \mathrm{d}u$$

where:

- u is a function that simplifies with differentiation
- $\mathrm{d}v$ is something that can be integrated

$$\begin{aligned} u &= w & \mathrm{d}v &= w \cdot e^{-\frac{k}{2}w^2}\mathrm{d}w \\ \implies \mathrm{d}u &= \mathrm{d}w & \implies v &= \int w \cdot e^{\frac{k}{2}w^2}\mathrm{d}w \\ & & \implies v &= \tfrac{1}{k}e^{\frac{k}{2}w^2} \end{aligned}$$

Hence the value of the variance may be solved:

Now that the expected value has been confirmed, consider the variance of the distribution:

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 \times f(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 \times \left(\sqrt{\frac{k}{2\pi}}e^{-\frac{k}{2}(x-\mu)^2}\right)\mathrm{d}x \\ &= \sqrt{\frac{k}{2\pi}}\int_{-\infty}^{\infty} (x-\mu)^2 \times \left(e^{-\frac{k}{2}(x-\mu)^2}\right)\mathrm{d}x \end{aligned}$$

Now observe that $(x-\mu)$ appears as an exponential and as a factor if this is redefined as $w = x - \mu \implies \mathrm{d}x = \mathrm{d}w$ we have:

$$\sigma^2 = \sqrt{\frac{k}{2}}\int_{-\infty}^{\infty} w^2 e^{-\frac{k}{2}w^2}\mathrm{d}w$$

Now the integrand is of the form $f(x) \times g(x)$ meaning that the only strategy to potentially deal with it is integration by parts:

$$\int u \mathrm{d}v = u \cdot v - \int v \mathrm{d}u$$

where:

- u is a function that simplifies with differentiation
- $\mathrm{d}v$ is something that can be integrated

$$\begin{aligned} u &= w & \mathrm{d}v &= w \cdot e^{-\frac{k}{2}w^2}\mathrm{d}w \\ \implies \mathrm{d}u &= \mathrm{d}w & \implies v &= \int w \cdot e^{\frac{k}{2}w^2}\mathrm{d}w \\ & & \implies v &= \tfrac{1}{k}e^{\frac{k}{2}w^2} \end{aligned}$$

Hence the value of the variance may be solved:

$$\begin{aligned}\sigma^2 &= \sqrt{\frac{k}{2\pi}}\int_{-\infty}^{\infty} w^2 e^{-\frac{k}{2}w^2}\mathrm{d}w \\ &= \sqrt{\frac{k}{2\pi}}\left[u \cdot v - \int v\mathrm{d}u\right]_{\infty}^{\infty} \\ &= \sqrt{\frac{k}{2\pi}}\left(\left[\frac{-w}{k}\cdot e^{-\frac{k}{2}w^2}\right]_{\infty}^{\infty} - \frac{1}{k}\int_{-\infty}^{\infty} e^{\frac{k}{2}w^2}\mathrm{d}w\right) \\ &= \sqrt{\frac{k}{2\pi}}\left[\frac{-w}{k}\cdot e^{-\frac{k}{2}w^2}\right]_{\infty}^{\infty} - \frac{1}{k}\left(\sqrt{\frac{k}{2\pi}}\int_{-\infty}^{\infty} e^{\frac{k}{2}w^2}\mathrm{d}w\right)\end{aligned}$$

The left term evaluates to zero and the right term is the area beneath the bell curve with mean value 0 and so evaluates to 1:

$$\begin{aligned}\sigma^2 &= 0 - \frac{1}{k} \\ \implies k &= \frac{1}{\sigma^2}\end{aligned}$$

So the function for the density curve can be simplified:

$$\begin{aligned}&= -\sqrt{\frac{k}{2\pi}}\cdot e^{k\cdot\frac{(x-\mu)^2}{2}} \\ &= \sqrt{\frac{1}{2\pi\sigma^2}}\cdot e^{\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}}\end{aligned}$$

now let $z = \frac{x-\mu}{\sigma} \implies \mathrm{d}z = \frac{\mathrm{d}x}{\sigma}$, this then simplifies to:

$$f(x) = \sqrt{\frac{1}{2\pi}}\cdot e^{-\frac{1}{2}z^2}$$

Now using the power series identity from BEFORE :

$$e^{-\frac{1}{2}z^2} = \sum_{n=0}^{\infty}\left[\frac{\left(-\frac{1}{2}z^2\right)^n}{n!}\right]$$

We can solve the integral of $f(x)$ (which has no elementary integral.

$$
\begin{aligned}
f\left(x\right) &= \sqrt{\frac{1}{2\pi}} \cdot \sum_{n=0}^{\infty}\left[\frac{\left(-\frac{1}{2}z^{2}\right)^{n}}{n!}\right] \\
\int f\left(x\right)\mathrm{d}x &= \frac{1}{\sqrt{2\pi}}\int\sum_{n=0}^{\infty}\left[\frac{\left(-\frac{1}{2}z^{2}\right)^{n}}{n!}\right]\mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \cdot \sum_{n=0}^{\infty}\left[\int\frac{\left(-1\right)^{-1}z^{2n}}{2^{n}\cdot n!}\mathrm{d}z\right] \\
&= \frac{1}{\sqrt{2\pi}} \cdot \sum_{n=0}^{\infty}\left[\frac{\left(-1\right)^{n}\cdot z^{2n+1}}{2^{n}\left(2n+1\right)n!}\right]
\end{aligned}
$$

Although this is a power series it still gives a method to solve the area beneath the curve of the density function of the normal distribution.

3 Understanding the p-value

Let's say that I'm given 100 vials of medication and in reality only 10 of them are actually effective.

POS	POS	POS	POS	POS	POS	POS	POS	POS	POS
:-:	:-:	:-:	:-:	:-:	:-:	:-:	:-:	:-:	:-:
NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG
NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG
NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG
NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG
NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG
NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG
NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG
NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG
NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG
NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG	NEG

We don't know which ones are effective so It is necessary for the effective medications to be detected by experiment. Let:

- the p-value be 9% for detecting a significant effect
- assume the statistical power is 70%

So this means that the corresponding errors are:

1. Of the 90 Negative Drugs, $\alpha \times 90 \approx 8$ will be identified as Positive (False Positive) a. This means 72 will be correctly identified as negative. (TN)
2. Of the 10 Good drugs $\beta \times 10 = 3$ will be labelled as negative (False Negative) b. This means 8 will be correctly identified as positive (True Positive)

These results can be summarised as:

	Really Negative	Really Positive
Predicted Negative	TNR; $(1 - \alpha)$	FNR; $\beta \times 10 = 3$
Predicted Positive	FPR; $\mathsf{FPR} = \alpha \times 90 \approx 8$	TPR $(1 - \beta)$

And a table visualising the results:

TP	TP	TP	TP	TP	TP	TP	FN	FN	FN
:-:	:-:	:-:	:-:	:-:	:-:	:-:	:-:	:-:	:-:
FP	FP	FP	FP	FP	FP	FP	FP	TN	TN
TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
TN	TN	TN	TN	TN	TN	TN	TN	TN	TN

So looking at this table, it should be clear that:

- If the null hypothesis had been true, the probability of a False Positive

would indeed have been $\frac{8}{90} \approx 0.09$

- The probability of incorrectly rejecting the null hypothesis though is the

number of FP from anything identified as positive $\frac{7}{7+6} \approx 0.5$

3.1 False Positive Rate

The False Positive Rate is expected to be α it is:

$$\begin{aligned}
\mathrm{E}\left(\mathsf{FPR}\right) &= \alpha; \\
\mathsf{FPR} &= \frac{FP}{N} \\
&= \frac{FP}{FN + TP} \\
&= \frac{8}{8 + 72} \\
&= 9\%
\end{aligned}$$

3.2 False Discovery Rate

The False discovery Rate is the proportion of observations considered as positive (or significant) that are False Positives. If you took all the results you considered as positive and pulled one out, the probability that one was a false positive (and you were commiting a type I error) would be the FDR and could be much higher than the FPR.

3.3 Measuring Probability

In setting α as 9% I've said that 'if the null hypothesis was true and every vial was negative, 9% of them would be false positives', this means that in practice 9% of the negative vials would be detected as false positives (I wouldn't count the positives because my α assumption was made under the assumption that everything was negative, hence 9% of the negative vials will be false positives).

So this measures the probability of rejecting the null hypothesis if it were true.

It does not measure the probability of rejecting the null hypothesis but then being mistaken, because to reject the null hypothesis it is necessary to consider observations that are considered positive (whether or not they actually are), the number of those that are False Positive would represent the probability of committing a type 1 error in that experiment

So the p -value measures the probability of committing a type I error under the assumption that the null hypothesis is true.

The FDR represents the actual probability of committing a type I error when taking multiple comparisons.

3.4 Comparing α and the p-value

The distinction between α and $p$ -value is essentially that the α value is set as a significance standard and the $p$ -value represents the probability of getting a test-statistic $\geq$ the observed value

The α value is the probability of

> Rejecting the null hypothesis under the assumption that the null hypothesis is true.

This will be the False Positive Rate:

> The proportion of Negative Observations misclassified as Positive will be the False Positive Rate.

Be careful though because this is not necessarily the *probability of incorrectly rejecting the null hypothesis* there is also the the $\mathsf{FDR} = \frac{\mathsf{FP}}{\mathsf{TP+FP}}$:

> The proportion of observations classified as positive that are false positives, this estimates the probability of rejecting the null hypothesis and being wrong. (whereas the α value is the probability of rejecting the null hypothesis under the assumption it was true this is different from the probability of rejecting H_0 and being wrong, which is the FDR).

The $p$ -value is the corresponding probability of the test statistic that was returned, so they mean essentially the same thing, but the α value is set before hand and the $p$-value is set after the fact:

> The $p$-value is the probability, under the assumption that there is no true effect or no true difference, of collecting data that shows a difference equal to or more extreme than what was actually observed.

3.5 Wikipedia Links

Helpful Wikipedia Links

- False Positive Rate
- False Discovery Rate
- Sensitivity and Specificity
- ROC Curve

  - This has all the TP FP calculations
- Type I and Type II Errors
  - This has the useful Tables and SVG Density Curve

4 Calculating Power

Statistical Power is the probability of rejecting the null hypothesis assuming that the null hypothesis is false (True Positive).

Complementary to the *False Positive Rate* and *False Detection Rate*, the power is distinct from the probability of correctly rejecting the null hypothesis, which is the probability of selecting a True Positive from all observations determined to be positive (the Positive Predictive Value or the Precision):

$$
\begin{aligned}
PPV &= \frac{TP}{TP + FP} \\
FDR &= \frac{FP}{TP + FN} \\
\alpha = \frac{FP}{N} &= \frac{TP}{TN + FP} \\
\beta = \frac{TP}{P} &= \frac{TP}{TP + FN}
\end{aligned}
$$

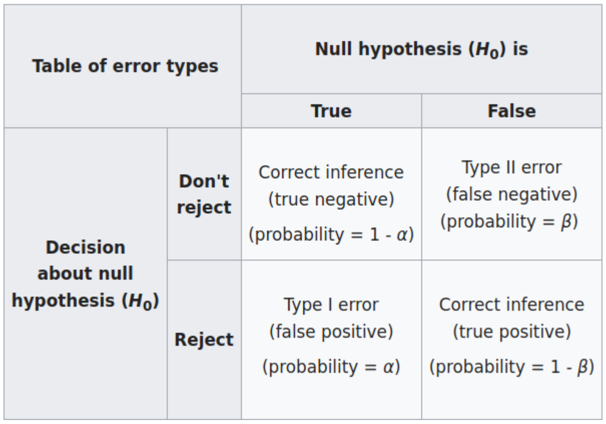


Table of error types		Null hypothesis (H_0) is	
		True	False
Decision about null hypothesis (H_0)	Don't reject	Correct inference (true negative) (probability = 1 - α)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = 1 - β)

4.1 Example

4.1.1 Problem

An ISP stated that users average 10 hours a week of internet usage, it is already known that the standard deviation of this population is 5.2 hours. A sample of $n = 100$ was taken to verify this claim with an average of $\bar{x}$.

A worldwide census determined that the average is in fact 12 hours a week not 10.

4.1.2 Solution

1. Hypotheses
   (a) H_0 : **The Null Hypothesis** that the average internet usage is 10 hours per week
   (b) H_a : **The Alternative Hypothesis** that the average internet usage exceeds 10 hours a week
2. Data

Value	Description
$n = 100$	The Sample Size
$\sigma = 5.2$	The Standard Deviation of internet usage of the population
$\mu = 10$	The alleged average internet usage.
$\bar{x} = 11$	The average of the sample
$\mu_T rue = 12$	The actual average from the population
$\alpha = 0.05$	The probability of a type 1 error at which the null hypothesis is rejected
$\beta = ??$	The probability of a type 2 error

4.1.3 Step 1; Find the Critical Sample Mean ($\bar{x}_{\text{crit}}$)

The Central Limit Theorem provides that the mean value of a sample that is:

- sufficiently large, or
- drawn from a normally distributed population

will be normally distributed, so if we took various samples of a population and recorded all the sample means in a set $\overline{X}$ we would have:

$$\overline{X} \sim \mathcal{N}\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)\right)$$

And hence we may conclude that:

$$\begin{aligned} Z &= \frac{\overline{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \\ \implies \overline{x}_{crit} &= \mu + z_{\alpha} \cdot \left(\frac{\sigma}{\sqrt{n}}\right) \\ \overline{x}_{crit} &= \mu + z_{0.05} \cdot \left(\frac{\sigma}{\sqrt{n}}\right) \\ \overline{x}_{crit} &= \mu + 1.645 \cdot \left(\frac{5.2}{\sqrt{100}}\right) \\ &= 10.8554 \end{aligned}$$

Thus H_0 is rejected for a sample mean of 10.86 hours per week at a confidence level of $\alpha = 0.05$.

4.1.4 Step 2: Find the Difference between the Critical and True Means as a Z-Value (prob of Type II)

The probability of accepting the null hypothesis assuming that it is false, is the probability of getting a value less than the critical value given that the mean value is actually 12:

$$\begin{aligned} z &= \frac{\overline{x}_{crit} - \mu_{true}}{\left(\frac{\sigma}{\sqrt{n}}\right)} \\ &= \frac{10.86 - 12}{\frac{5.2}{10}} \qquad = -2.2 \end{aligned}$$

4.1.5 Step 3: State the value of β

$$\begin{aligned} \beta &= \mathrm{P}\left(\textsf{Type II Error}\right) \\ &= \mathrm{P}\left(H_0 \textsf{ is not rejected} \mid H_0 \textsf{ is false}\right) \\ &= \mathrm{P}\left(\mu_{\overline{X}_{\textsf{Crit}}} < \overline{x}_{\textsf{crit}} \mid \mu = 12\right) \\ &= 0.014 \end{aligned}$$

4.1.6 Step 4: State the Power Value

$$\begin{aligned}
\textsf{Power} &= (H_0 \textsf{ is not rejected} \mid H_0 \textsf{ is false})\\
&= \mathrm{P}\left(\mu_{\overline{X}_{\textsf{Crit}}} < \overline{x}_{\textsf{Crit}}\right)\\
&= 1 - \beta\\
&= 1 - 0.14\\
&= 98.6\%
\end{aligned}$$

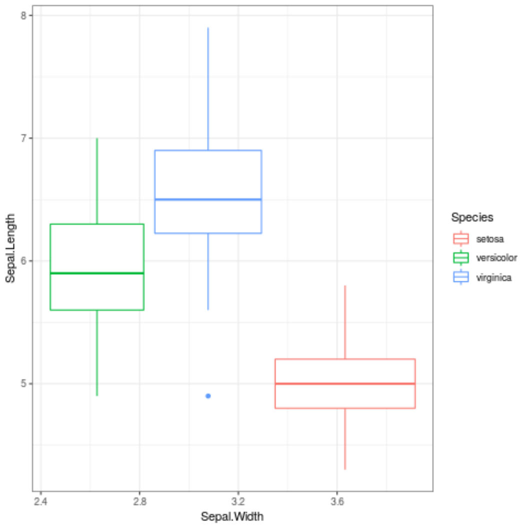
5 Weekly Material

5.1 Wk 3; Smoking and Birth Weight

5.1.1 Boxplots

The delimiting marks in box plots correspond to the median and interquarlite range (which is basically the median of all data below the median):

```
library("ggplot2")
ggplot(iris, aes(x = Sepal.Width, y = Sepal.Length, color = Species)) +
geom_boxplot() +
theme_bw()
```



5.1.2 Lecture Announcements

Everything is online now. We'll be using *Zoom* a lot.

1. **TODO** Finish Quiz 1 30 minutes to finish it, Test your computer First.

2. Post `pacman` on the mailing list.

5.1.3 Naming Variables

Attribute . . . Data Base

5.1.4 TODO Review Chi Distribution,

1. Is it in VNote?

2. Should I put it in `org-mode`?

5.1.5 TODO Fix YAML Headers in `rmd` to play ball with Notable

1. **TODO** Post this use-case to Reddit

2. **TODO** Fix YAMLTags and TagFilter and Post to Reddit

3. **DONE** Is there a way to fix the Text Size of Code in emacs when I zoom out? Yeah just disable `M-x mixed-pitch-mode`

5.1.6 Calculalating mean

```r
library(tidyverse)
bwt <- c(3429, 3229, 3657, 3514, 3086, 3886)
(bwt <- sort(bwt))
mean(bwt)
mean(c(3429, 3514))
median(bwt)
max(bwt)-min(bwt)
```

The mean value is nice in that it has good mathematical properties, so for predictions and classifications (like gradient descent), if the model contains the mean the model will be smooth and the mean will lead to a well behaved model with respect to the derivative.

The Median value, however is more immune to large outliers, for example:

```r
library(tidyverse)
x <- c(rnorm(10), 9) * 10 %>% round(1)
mean(x); median(x)
```

5.1.7 Calculating Range

```
range(bwt)
bwt %>% range %>% diff
```

5.1.8 Calculating Variance

```
(var <- (bwt-mean(bwt))^2 %>%  mean)
var(bwt)
(sd <- (bwt-mean(bwt))^2 %>%  %>% sqrt) # Not using n-1 !!
(sd <- sqrt(sum((bwt-mean(bwt))^2)/(length(bwt) -1)))
sd(bwt)
mean(sum((bwt-mean(bwt))^2))
```

5.1.9 InterQuartile Data

6 Central Limit Theorem

The central Limit theorem provides us the sampling distribution of $\overline{X}$ even when we don't know what the original population of X looks like:

1. If the population is normal, the sample mean of that population will be

normally distributed, $\overline{X} \sim \mathcal{N}\left(\mu\left(\frac{\sigma}{\sqrt{n}}\right)\right)$

1. As sample size n increases, the distribution of sample means converges to

the population mean μ

- i.e. the *standard error of the mean*

$\sigma_{\overline{x}} = \left(\frac{\sigma}{\sqrt{n}}\right)$ will become smaller

1. If the sample size (of sample means) is large enough ($n \geq 30$) the sample

means will be normally distributed even if the original population is non-normal