Thinking About Data

Ryan G

April 20, 2020

Contents

- PDF Version
- HTML Version

1 TODO Overheads

1.0.1 Gitignore Quiz

Put the Quizz in a .gitignore file here Clear the bad lines

1.0.2 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.

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

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

1.0.4 TODO Fix MkDocs for New 'MD' Structure

1.0.5 TODO Put all RMD HTML Files on a GitPage

Look at Using Bookdown rather than mkdocs for this?

2 Unit Information

- Learning Guide
  - Zoom Tutorial
  - Zoom Lecture

3 Deriving the Normal Distribution

3.1 Power Series SERIES

A function f :

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

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

Is a Power Series a and will either:

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

3.1.1 Example

Take some function equal to the following power series:

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

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 $x = 0$.

3.1.2 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. 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:

[1] Refer to Solving Series Strategy

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

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

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

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

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

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

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

3.1.3 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

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

3.1.4 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 $f(z) = \sum_{n=1}^{\infty} c_n (z-a)^n$:

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

Following this pattern forward:

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

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

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

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

1. Power Series Expansion of e

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

3.2 Modelling Normal Distribution

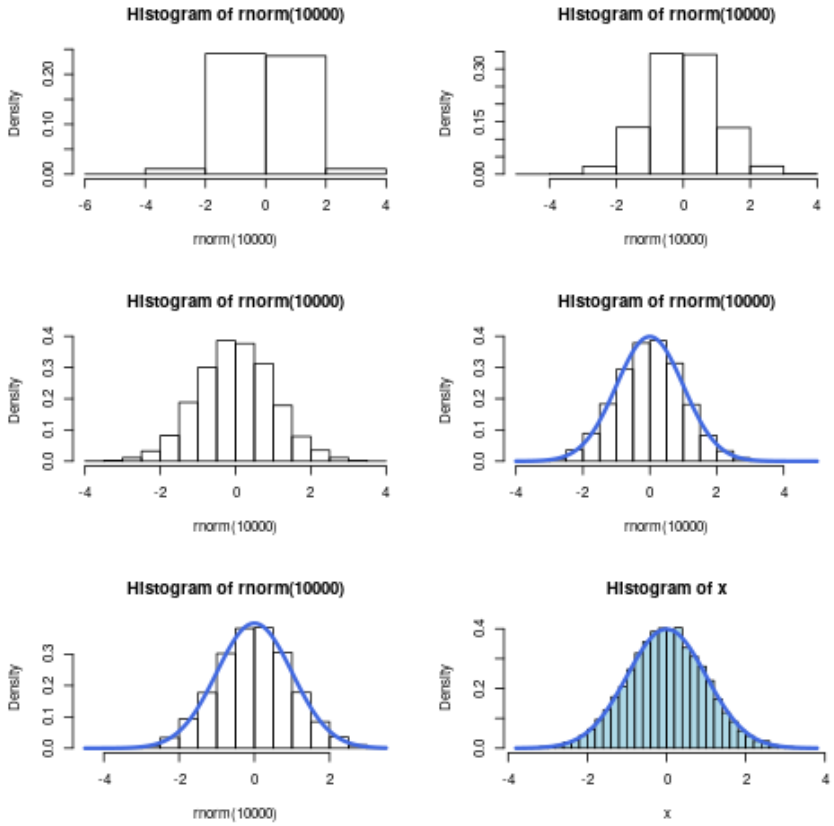
The Normal Distribution is a probability density function that is essentially modelled after observation.[2]

3.2.1 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))
 layout(matrix(1:6, 3, 2, byrow = TRUE))


 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) as described in listing 1 and shown in figure **??**

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

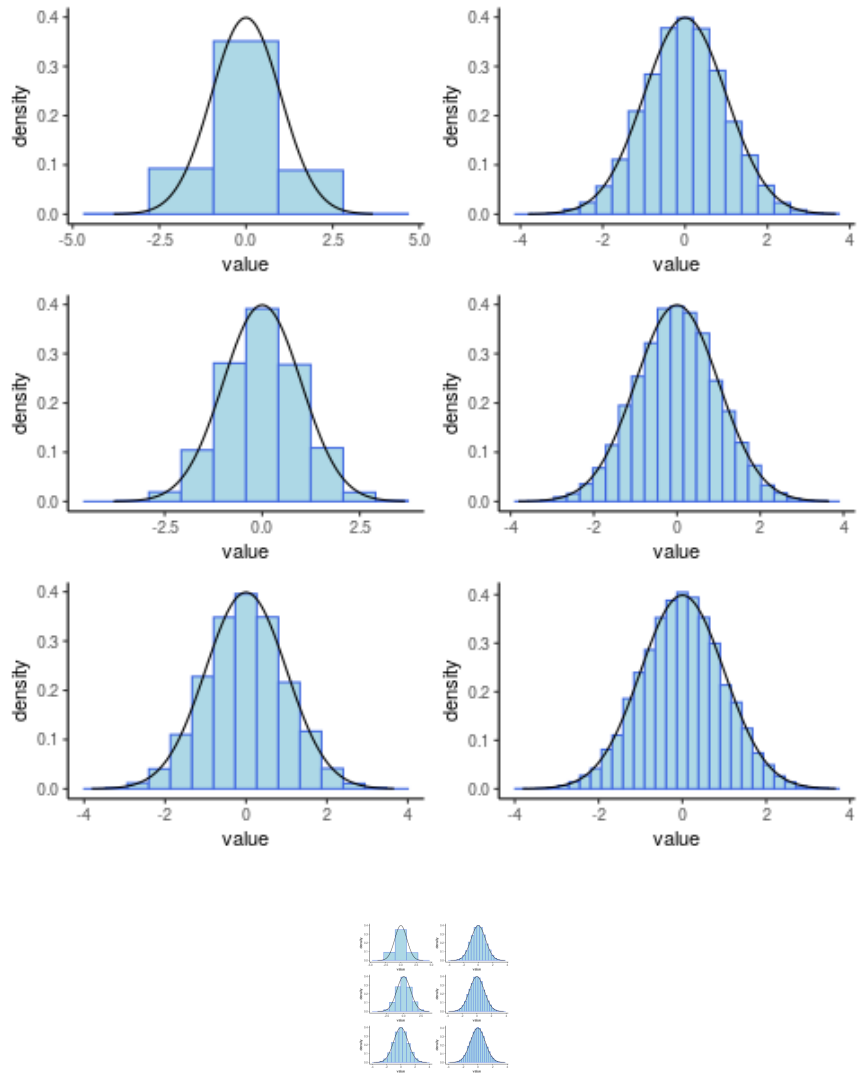


Figure 1: Histograms Generated in ggplot2

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.

3.2.2 Defining the Normal Distribution

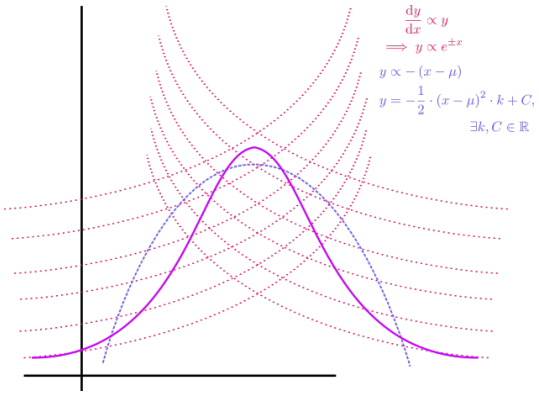
Data are said to be normally distributed if, the plot of the frequency density curve is such that:

- 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 3.2.3, 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 figure **??**:

Bell Curve



3.2.3 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:

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

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

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

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 $\frac{\mathrm{d}f}{\mathrm{d}x} \propto -(x-\mu)$

3.2.4 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 $\frac{\mathrm{d}f}{\mathrm{d}x} \propto f$:

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

3.2.5 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:

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

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:

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

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:

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

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:

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

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 $f(x) \times g(x)$ 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:

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.

4 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 $\frac{7}{7+6} \approx 0.5$

4.1 False Positive Rate

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

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

4.2 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.

4.3 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.

4.4 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 $\mathsf{FDR} = \frac{\mathsf{FP}}{\mathsf{TP}+\mathsf{FP}}$:

> 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.

4.5 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

5 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):

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

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)

5.1 Example

5.1.1 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.

5.1.2 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

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

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 $\alpha = 0.05$.

5.1.4 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:

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

5.1.5 Step 3: State the value of β

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

5.1.6 Step 4: State the Power Value

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

6 Weekly Material

6.1 (3); Comparison of Population Samples WK3

6.1.1 Lecture

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()
```

2. Lecture Announcements Everything is online now. We'll be using *Zoom* a lot.
   (a) **DONE** Finish Quiz 1 30 minutes to finish it, Test your computer First.
   (b) Post `pacman` on the mailing list.
3. Naming Variables Attribute ... Data Base
4. **DONE** Review Chi Distribution,
   (a) Is it in VNote?
   (b) Should I put it in `org-mode`?
5. **DONE** Fix YAML Headers in `rmd` to play ball with Notable
   (a) **DONE** Post this use-case to Reddit

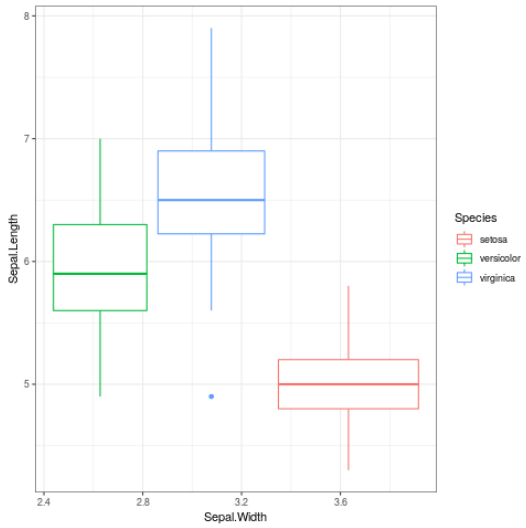


Figure 2: Bar Plots Generated in ggplot2

(b) **DONE** Fix YAMLTags and TagFilter and Post to Reddit BASH
The Bash Script is here

i. Should I have all the lists and shit in `/tmp`

Pros	Cons
Less Mess	Harder to Directly watch what's happening
Easier to manage [5]	

should I use `/tmp` or `/tmp/Notes` or somting?

ii. **DONE** Is there an easy way to pass the md off to vnote Or should I just use the `ln -s ~/Notes/DataSci /tmp/notes` trick?
Follow the instructions here, it has to be done manually and then symlinked SCHEDULED: *<2020-03-21 Sat>*

iii. Should I have all the lists and shit in `/tmp`

[5] By which I mean I'm not sure if the directory that the `00tagmatch` directory `00taglist` file will be made in are the wd of bash, or, if they are the location of `~/Notes`
I'm also not sure how that will be influenced by looking for `#tags` in the `~/Notes/DataSci` Directory

Pros	Cons
Less Mess	Harder to Directly watch what's happening
Easier to manage [5]	

should I use `/tmp` or `/tmp/Notes` or somting?

(c) **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`

6. Calculalating mean

```
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:

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

7. Calculating Range

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

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))
```

9. InterQuartile Data

6.1.2 Tutorial

The tutorial work is located at `~/Notes/DataSci/ThinkingAboutData` and linked here:

- PDF
- HTML
- MD
- RMD

6.2 DONE (4); Using Student's t-Distribution WK4

6.2.1 Material

- Lecture 04
- Worksheet 04

6.2.2 DONE Lecture

1. The Wilcoxon-Mann-Whitney Test Often used when the data is skewed and the use of the t -test is suspect. :ID: 3bcc3e01-598c-484d-859f-8ca82fe340e6 :DIR: Attachments/Thinking$_{\text{AboutData}}$/

```
bwt = c(3429, 2329, 3657, 3514, 3086, 3886)
smoke = c(0, 0, 1, 0, 1, 0)
prob = data.frame(bwt, smoke)

# How to actually do that for the ordinary data

wilcox.test(bwt ~ smoke, birthwt)
```

2. *Student*'s t-test Consider the Histograms from the birthweight data shown below in figure 4 as generated by listing 4; this data represents a sample from two populations:

    - Those that Smoked During Pregnancy

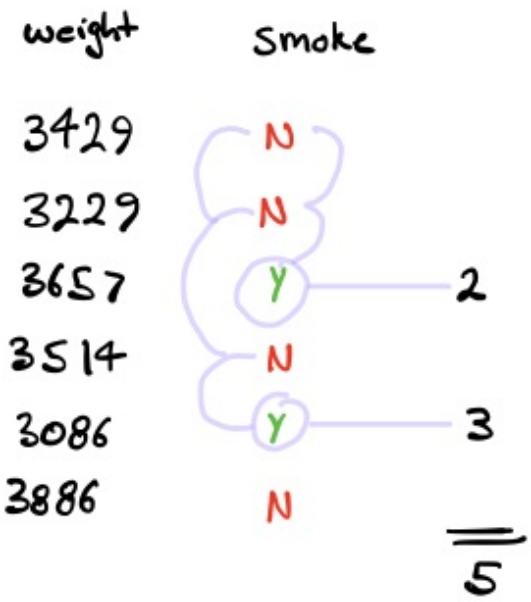


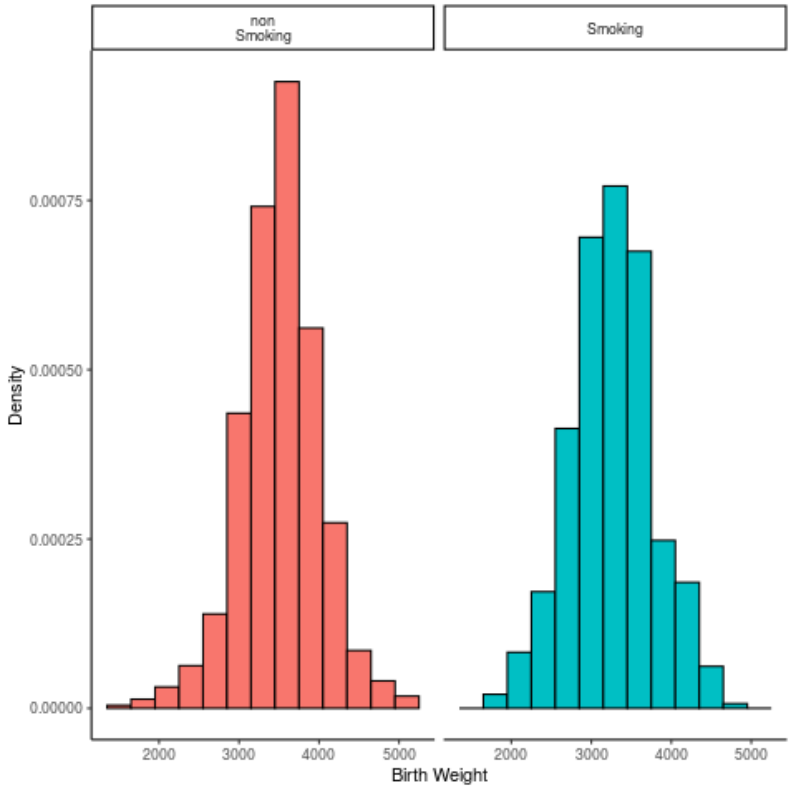
Figure 3: Attempt at man whitney test

- Those that did not smoke During Pregnancy

```
load.pac <- function() {

  if(require("pacman")){
    library(pacman)
  }else{
    install.packages("pacman")
    library(pacman)
  }

  pacman::p_load(xts, sp, gstat, ggplot2, rmarkdown, reshape2, ggmap,
                 parallel, dplyr, plotly, tidyverse, reticulate, UsingR, Rmpf
                 swirl, corrplot, gridExtra, mise, latex2exp, tree, rpart,
                 lattice,
                 rstudioapi)

}

load.pac()
```

```
birthwt <- read.csv("./Attachments/Thinking_About_Data/birthwt.csv")
birthwt_pretty <- birthwt
birthwt_pretty$smoke <- ifelse(birthwt_pretty$smoke=="yes", TRUE, FALSE)
birthwt_pretty$smoke <- ifelse(birthwt_pretty$smoke, "Smoking", "non\nSmoking

hist <- ggplot(birthwt_pretty, aes(x = bwt, fill = smoke, col = "black", y =
         theme_classic() +
         labs(x = "Birth Weight", y = "Density") +
         geom_histogram(binwidth = 300, col = "black") +
         facet_grid(. ~ smoke) +
         guides(fill = FALSE)
hist
```

Figure 4: Using GGplot two create a facet grid of histograms



The whole idea is that these two distributions have a similar shape and

can be related by just standardising them relative to 1 and 0 by sliding left and right (addition) and scaling by (multiplication).

3. Central Limit Theorem Now if we had a different sample from the population we would have a different sample mean value ($\bar{x}$ would change even though μ would be the same)

   what we need to become interested in is the distribution $\bar{x}$[6].

   The Central Limit Theorem (which was considered whenCalculating Power) essentially provides that the distribution of means (the sampling distribution denoted $\bar{X}$) will follow a normal distribution centred around μ and with a spread inversely proportional to the root of the sample size:

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

   (a) What is Standardisation This means that all distributions can be transformed relative to the standard normal distribution and inferences relating to the area under the curve (the p -value usually) can be read off from tables:

$$z_i = \frac{x_i - \bar{x}}{s}$$

   (b) Standard Error The standard error of the mean is the standard deviation of sampling mean.

   (c) An Example

   If we had a population, for the sake of argument say:

   - $\mu = 7$
   - $\sigma = 13$

   It would have a Histogram that would look something like figure 6:

```
library(latex2exp)
library(tidyverse)

hist(rnorm(99999, 7, 13), main = "Population of Observations with mu = 7, s
```

[6] although we could talk about other sampling statistics like the sampling standard deviation or the median, for this re-sampling is used, usually the bootstrap method (Source)

```
mu  <- 7
sig <- 13
x <- rnorm(999999, mu, sig)
data <- tibble("obs" = x)

ggplot(data, aes(x = obs)) +
  geom_histogram(aes( y = ..density..), fill = "lightblue", bins = 50) +
  stat_function(fun = dnorm,
                args = list(mean = mu, sd = sig),
                col = "orchid", lwd = 1.5) +
  theme_bw() +
  labs(x = "Observation Value", title = TeX("Simulated Population with $\\mu
```

Figure 5: Simulated Population

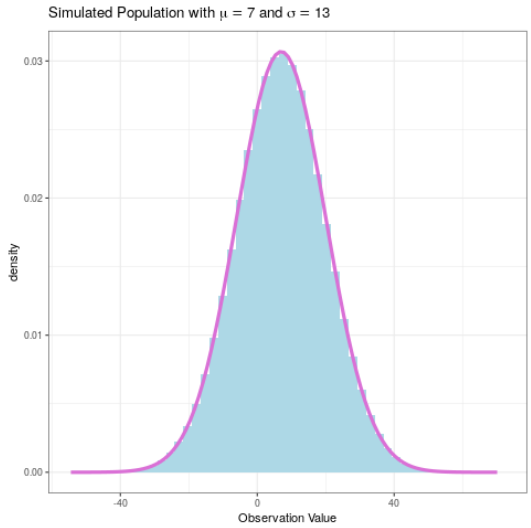


Figure 6: Population Histogram Generated in ggplot2

If samples of size n = 100 were taken from this population many times (say 999 times) the distribution of the mean value would look like figure 7:

```
library(tidyverse)

vals <- replicate(999, {rnorm(100, 7, 13) %>% mean()})
```

```
hist(vals, freq = FALSE)
sd(vals)
```

Figure 7: Observed sampling Distribution

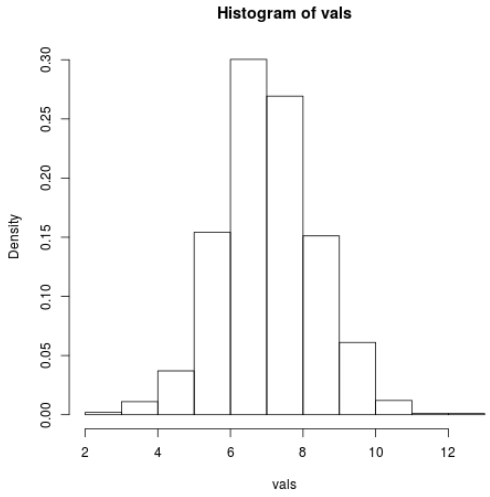


Figure 8: Sampling Distribution Generated in ggplot2

And the corresponding standard deviation and mean values would be:

```
print(paste("mean value of samples: ", signif(mean(vals),3)))
print(paste(" std. dev. of samples: ", signif(sd(vals),3)))
print(paste(" sd/sqrt(n): ", signif(13/sqrt(100),3)))
```

So the distribution of a mean value from a population will be normally distributed around the population mean with a standard deviation of $\frac{\sigma}{\sqrt{n}}$.

(d) Why is the Std. Error $\frac{\sigma}{\sqrt{n}}$ I should investigate this

4. Standard Error for a Difference in Means I made some markdown notes on this here below is the conversion:

   (a) Standard Error Generally If we had a sample of size n from a population [1] and considered the mean value, that mean value itself would belong to:

- the population of all possible sample mean values

Bit of a mouthfull, but the idea being that if the sample was repeated the set of all mean values $\bar{X}$ mean value would follow a distribution:

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

and the spread of that distribution would be known as the standard error.

So from this there are naturally two things to consider:

i. The Standard Error for the difference between two population means
   A. i.e. what's the standard error for $\mu_1 - \mu_2$ given the sample of both populations that was taken?

ii. What is the sample standard deviation of both groups put together (s_p?
   A. Generally if the populations are expected to have the same spread this *pooled* std deviation can be assigned to either group as a sort of average.

(b) Standard Error of the Difference The deviation of the difference will be the deviation of the first term plus the deviation of the second term:

$$\begin{aligned}
\operatorname{sd}(\overline{x}_1) = \sigma_{\overline{x}_1} = \frac{\sigma_1}{\sqrt{n}}; \qquad \operatorname{var}(\overline{x}_1) = \sigma^2_{\overline{x}_1} = \frac{\sigma_1^2}{n} \\
\operatorname{sd}(\overline{x}_2) = \sigma_{\overline{x}_2} = \frac{\sigma_2}{\sqrt{n}}; \qquad \operatorname{var}(\overline{x}_2) = \sigma^2_{\overline{x}_2} = \frac{\sigma_2^2}{n}
\end{aligned}$$

Adding/Subtracting the values will mean that the variance will also add:

$$\begin{aligned}
\operatorname{var}(\overline{x}_1 + \overline{x}_2) = \sigma^2_{\overline{x}_1} + \sigma^2_{\overline{x}_1} = \sigma^2_{\overline{x}_1 \pm \overline{x}_2} \\
\operatorname{sd}(\overline{x}_1 + \overline{x}_2) = \sqrt{\sigma^2_{\overline{x}_1} + \sigma^2_{\overline{x}_1}} = \sigma_{\overline{x}_1 \pm \overline{x}_2}
\end{aligned}$$

Now remember that the standard error is:

$$\operatorname{SE}\left(\overline{X}\right) = \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

Hence:

$$\begin{aligned}
\mathrm{sd}\left(\overline{x}_1 + \overline{x}_2\right) = \sigma_{\overline{x}_1 \pm \overline{x}_2} &= \sqrt{\sigma_{\overline{x}_1}^2 + \sigma_{\overline{x}_1}^2} \\
&= \sqrt{\left(\sigma_{\overline{x}_1}\right)^2 + \left(\sigma_{\overline{x}_1}\right)^2} \\
&= \sqrt{\left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2} \\
&= \sqrt{\left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2} \\
&= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
\end{aligned}$$

Now presume that both populations have the same standard deviation, this means that the standard deviation of all the observations **p*ooled* together will be the same (or in the case of sample standard deviation be a better predictor):

$$\begin{aligned}
\sigma_{\overline{x}_1 \pm \overline{x}_2} &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\
\sigma_{\overline{x}_1 \pm \overline{x}_2} &= \sqrt{\frac{\sigma_p^2}{n_1} + \frac{\sigma_p^2}{n_2}} \\
&= \sqrt{\sigma_p^2 \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}
\end{aligned}$$

But then it's fairly straight forward:

$$\begin{aligned}
\sigma_{\overline{x}_1 \pm \overline{x}_2} &= \sqrt{\sigma_p^2 \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \\
&= \sigma_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}
\end{aligned}$$

However we won't know the actual population standard deviation and so instead the sample standard deviation may be used as a predictor:

$$E\left(s_{\overline{x}_1 \pm \overline{x}_2}\right) = \sigma_{\overline{x}_1 \pm \overline{x}_2}$$

$$\implies s_{\overline{x}_1 \pm \overline{x}_2} = s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

(c) Pooled Standard Deviation The sample variance of both populations would be:

$$s_p = \sqrt{\frac{1}{n_1 - 1 + n_2 - 1} \cdot \sum_{i=1}^{n_1+n_2} \left[(x_i - \overline{x})^2\right]}$$

$$s_p^2 = \frac{1}{n_1 - 1 + n_2 - 1} \cdot \sum_{i=1}^{n_1+n_2} \left[(x_i - \overline{x})^2\right]$$

$$(n_1 - 1 + n_2 - 1) \cdot s_p^2 = \sum_{i=1}^{n_1+n_2} \left[(x_i - \overline{x})^2\right]$$

Now by splitting apart the sums and rewriting in terms of the sample standard deviation s:

$$(n_1 - 1 + n_2 - 1) \cdot s_p^2 = \sum_{j=1}^{n_1} \left[\left(x_{1_j} - \overline{x}_1\right)^2\right] + \sum_{k=1}^{n_1} \left[\left(x_{2_k} - \overline{x}_1\right)^2\right]$$

$$(n_1 - 1 + n_2 - 1) \cdot s_p^2 = (n_1 - 1) \times s_1 + (n_2 - 1) \times s_2$$

$$s_p^2 = \frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 - 1 + n_2 - 1}$$

$$s_p = \sqrt{\frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 - 1 + n_2 - 1}}$$

(d) The t-statistic
This difference in means can hence be standardised:

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\frac{s}{\sqrt{n}}} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
$$= \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

The $t-$ distribution is a type of distribution that, while shaped like a bell curve, is not actually normal, but it's close to normal, in fact the t distribution may be made arbitrarily close to a normal curve if the the sample size (or rather the *degrees of freedom*) is made sufficiently large.

The degrees of freedom is usually something to the effect of $n_1 + n_2 - 2$.

There's an example of a pen-paper solution on page 20 of the lecture slides.

5. Confidence Intervals A confidence interval is the expected range of sample means from the population given the observation.

   There's a distinction between 10 samples and 1 sample of 10 observations, be careful.

   So if the population was continuously resampled, what interval of values could be taken such that our observation would be a false positive only 5% of the time?

   This is made more confusing when talking about a difference in population means, because in that case the actual difference value is the sampled statistic.

   (a) What's the difference between Confidence and Prediction Confidence intervals are concerned with predicting the population mean value within an interval that corresponds to p-value, i.e. the mean value of the population lies within this interval with a 95% probability of incorrectly rejecting the null hypothesis under the assumption that it was actually true.

6. Hypothesis Test First you need to define your population, which is all observations that you could be interested. To a degree it is somewhat arbitrary.

   If you're population is really small it's probably not very interesting

(a) Null Hypothesis Usually nothing happened, what we desire to refute.

6.2.3 DONE Tutorial ATTACH

The File is here

6.3 DONE (5) Discrete Distributions (Mapping Disease)WK5

- Lecture
- Practical
  - RMD File

6.3.1 DONE Quizz for t Distribution Material

The Quiz has been Released

- 03 Tutorial

Tutorial 3

- 04 Tutorial
- Lecture 03
- Lecture 04

6.3.2 Lecture

The Poisson Model is the Binomial Model stretch towards its limits.

1. Combinatorics

   The Counting Formulas are:

selection	ordered	unordered
With Repetition	n^m	$\binom{m+n-1}{n}$
Without Repetition	$n_{(m)}$	$\binom{n}{m}$

   Where:

   - $\binom{n}{m} = \frac{m_(n)}{n!} = \frac{m!}{n!(m-n)!}$

- $n_{(m)} = \frac{n!}{(n-m)!}$
- $n! = n \times (n-1) \times (n-2) \times 2 \times 1$

2. Binomial Distribution

   A Binomial experiment requires the following conditions:

   (a) We have n independent events.

   (b) Each event has the same probability p of success.

   (c) We are interested in the number of successes from the n trials (referred to as `size` in ***R***.

   (d) The probability of k successes from the n trials is a Binomial distribution with

   probabilities:

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

   (a) Problem Hard drives have an annual failure rate of 10%, what is the probability of 2 hard drives failing after 3 years?

   This means that:

   - The number of repetitions is 3 (n = `size` = 3)
   - The statistic we are interested in is 2 (k = `x` = 2)

```
dbinom(x = 2, size = 3, prob = 0.1)

## For Hard Drives
    ## k/x....is the number of years
    ## n/ ....size is the number of failures
    ## p  ....is the probability of failure

## choose(n,k)*p^k*(1-p)^(n-k)
## dbinom(x = 2, size = 4, prob = p)
```

   So in this case there would only a 2% chance.

3. Poisson An interesting thing with the poisson distribution is that the mean value and the variance are both equal.

   The expected value is the limit that the mean value would approach if the sample was made arbitrarily large, the value is denoted λ

Poisson is French for fish so sometimes people call the distribution the *fishribution*.

4. Binomial and Poisson

   The Poisson distribution is derived from the Binomial distribution.

   If $(1-p)$ is close to 1 then $np(1-p) \approx np$, so for very large sample sizes we have the expected value equal to the variance, and a *Poisson* distribution.

   This has something to do with widely increasing the number of trials, like say if we hade an infinite number of trials with the probability of success in a given hour as 30%.

   For a binomial distribution there are a set number of trials, let's say 8 trials with a 20% probability of Success:

1	2	3	4	5	6	7	8
F	S	F	F	S	S	F	F

   In this case there are 3 successes, so let's set $k = 3$ and instead however the region was divided into smaller spaces ($n \to \infty$):

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X	S	X	X	S	S	F	F	X	S	X	X	S	S	F	F

   If we kept dividing this up we would have

   The poisson is the limit as we increase the number of trials but try to keep k constant??

5. Confidence Intervals

   (a) Binomial Just use bootstrapping, but also you can just use an approximate standardisation:

$$\begin{aligned} \sigma &= \sqrt{p\,(1-p)} \\ \sigma_{\overline{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{\sqrt{p\,(1-p)}}{\sqrt{n}} \\ \implies p - z_{0.025} \times \sigma_{\overline{x}} < & p < p - z_{0.975} \times \sigma_{\overline{x}} \\ \implies p - 1.96\sigma_{\overline{x}} < & p < p - z_{0.975} 1.96\sigma_{\overline{x}} \end{aligned}$$

So an approximate 95% confidence interval could be

This is the normal data, but remember that this is estimating binomial by standard normal and so will only be good for large values of n because binomial is discrete by nature.

(b) Poisson This can also be done with Poisson by bootstrapping or using the same trick of λ as the variance:

$$\begin{aligned} \sigma &= \sqrt{\lambda} \\ \sigma_{\overline{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{\sqrt{\lambda}}{\sqrt{n}} \\ \implies \lambda - z_{0.025} \times \sigma_{\overline{x}} < & p < \lambda + z_{0.975} \times \sigma_{\overline{x}} \\ \lambda - 1.96\sigma_{\overline{x}} < & p < \lambda + 1.96\sigma_{\overline{x}} \\ p - 1.96\sqrt{\frac{\lambda}{n}} < & p < p - z_{0.975}1.96\sqrt{\frac{\lambda}{n}} \end{aligned}$$

6. Summary

- Binomial for independent trials
  - mean: np
  - variance: np(1-p)
    - standard error roughly is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Poisson for number of events in a given period
  - mean λ
  - variance λ
  - Standard error roughly is $\sqrt{\frac{\hat{\lambda}}{n}}$
- Choropleth maps are useful for visualising changes over area.

6.3.3 Tutorial

- 05_RMD File
- 05_MD File
- 05_PDF File
- 05_HTML File

6.4 DONE (6) Paired t-test (Observation or Experiment) WK6

- 06 Lecture Notes
- 06 Practical
  - RMD File

6.5 TODO (7) Corellation (Do Taller People Earn More) WK7

- 07 Lecture Notes
- 07 Practical
  - RMD File

6.5.1 DONE Lecture

In the past we did categorical and categorical-continuous.
Now we're doing purely continuous.

1. **DONE** How to Derive the Correlation Coefficient Refer to this paper
   The Correlation coefficient can be interpreted in one of two ways:
   - The covariance scaled relative to the x and y variance
     - $\rho = \frac{S_{x,y}}{s_x \cdot s_y}$
   - The rate of change of the line of best fit of the standardised data
     - This is equivalent to the rate of change of the line of best fit divided by $(\frac{s_y}{s_x})$:
       * $\rho = b \cdot \frac{s_x}{s_y} \quad \iff \quad \hat{y_i} = bx_i + c$

   This can be seen by performing linear regression in ***R***:

```
head(cars)
cars_std <- as.data.frame(scale(cars))
y <- cars$dist
x <- cars$speed

## Correlation Coefficient
cor(x = cars$speed, y = cars$dist)
```

```
  ## Covariance
  cov(x = cars$speed, y = cars$dist)/sd(cars$speed)/sd(cars$dist)

  ## Standardised Rate of Change
  lm(dist ~ speed, data = cars_std)$coefficients[2]

  ### Using Standardised Rate of change
  lm(dist ~ speed, data = cars)$coefficients[2] / (sd(y)/sd(x))
```

```
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
```

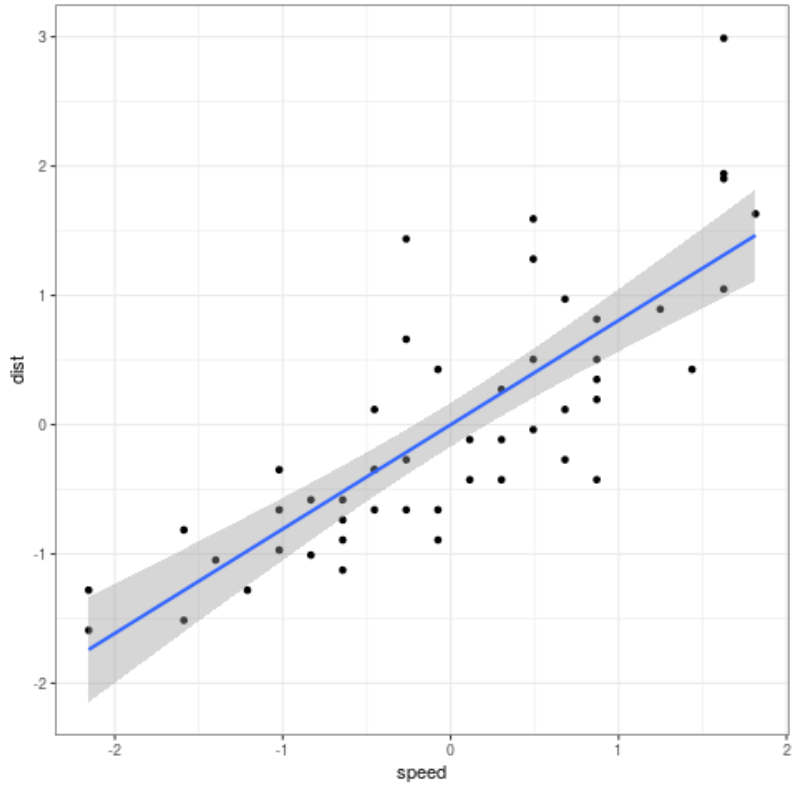
```
[1] 0.8068949
```

```
[1] 0.8068949
```

```
    speed
0.8068949
```

```
    speed
0.8068949
```

```
#+BEGIN_SRC R :cache yes :exports both :results output graphics :file ./test.png
  library("ggplot2")
  cars_std <- as.data.frame(scale(cars))

  ggplot(cars_std, aes(x = speed, y = dist)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_bw()
```



This shows that despite noise data can still have a correlation coefficient of 1 if the noise is evenly distributed in a way that the correlation coefficient can have a rate of change of 1.

2. **TODO** Prove the Correlation Coefficient and email Laurence

3. Bootstrapping The big assumption with bootstrapping is that the population can be seen as equiv to an infinite repetition of the sample size.

   So assume that a population that is an infinite repetition of the sample, then take a sample of that infinite population and you have a bootstrap.

   So we could either create a population from the sample with a size of ∞, which might be difficult, or, we could instead just resample the observation for each replication.

4. Confidence Intervals

```
load("Notes/DataSci/ThinkingAboutData/TAD.rdata ")
```

```
r = cor(crabsmolt$postsz, crabsmolt$presz)
a <- crabsmolt
N <- nrow(crabsmolt)
pos <- sample(N, size = N, replace = TRUE)
aboot <- a[pos,]

cor(aboot$postsz, aboot$presz)


#  replicate(10^4, {})
```

7 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