**Trabajo escrito sobre ANOVA de una vía y Kruskal-Wallis de una vía**

**Mejores épocas del año para visitar Basilea, Suiza para diferentes tipos de actividades turísticas.**

**Introducción**

Basilea es una ciudad cultural para sibaritas que con sus más de 40 museos concentra el mayor número de museos del país. Además, la ciudad cuenta con un hermoso centro histórico, arquitectura moderna y el río Rin que invita a dar largos paseos.

El emblema secreto de Basilea es el Rin. En verano, la vida gira en torno a él en todas sus vertientes. Amantes del sol, paseantes, estudiantes y gente de negocios animan sus orillas en la misma medida. Ideal para refrescarse o para dar un agradable paseo en barco.

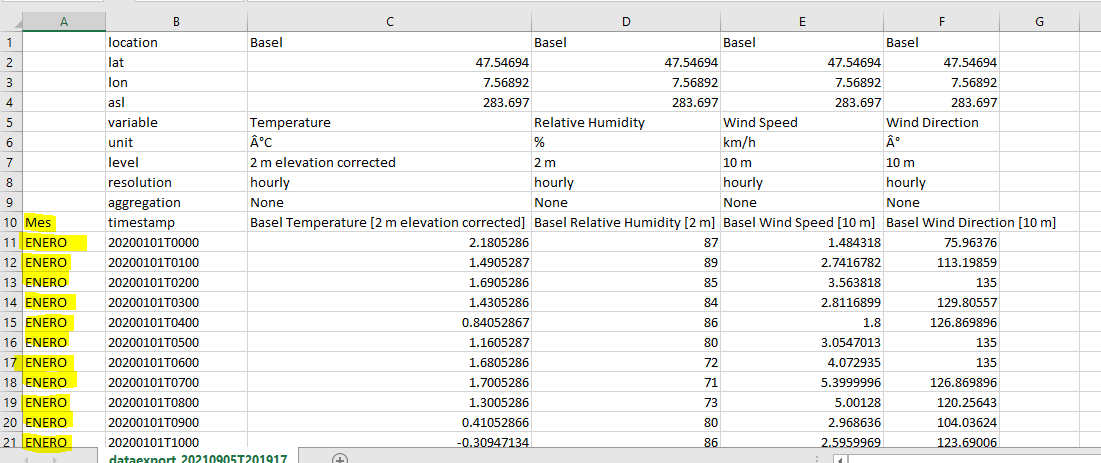
Mientras que en verano es cuando están la mayoría de los turistas ya que sus condiciones son ideales para practicar senderismo, ciclismo, kayak y disfrutar de otras actividades al aire libre; hay turistas que les gusta las actividades relacionadas a la nieve como andar en trineo, etc.

**Objetivos**

Este trabajo escrito lo utilizaremos para determinar en el año 2020 que meses son significativamente más fríos que otros en la ciudad de Basilea, Suiza con la finalidad de brindar información a los turistas sobre cuales son las mejores fechas para visitarlo en un futuro, dependiendo de las actividades que se quieran realizar. Este estudio se realizó con un alfa de 0.05. Para ello, nos valdremos de los datos recolectados por el servicio “meteoblue”. Se utilizó este link para obtener las temperaturas necesarias para las fechas <https://www.meteoblue.com/en/weather/archive/export/basel_switzerland_2661604?daterange=2020-01-01%20-%202020-12-31&domain=NEMSAUTO&params%5B%5D=temp2m&params%5B%5D=relhum2m&params%5B%5D=wind%2Bdir10m&min=2020-01-01&max=2021-01-01&utc_offset=2&timeResolution=hourly&temperatureunit=CELSIUS&velocityunit=KILOMETER_PER_HOUR&energyunit=watts&lengthunit=metric&degree_day_type=10%3B30&gddBase=10&gddLimit=30>.

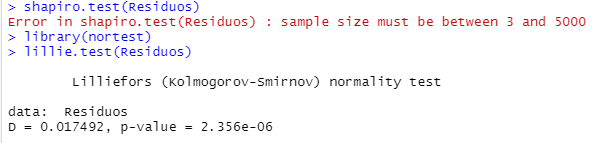
**Metodología estadística**

Lo primero que se debemos hacer es clasificar correctamente los datos a evaluar. El Excel que nos devuelve el servicio “meteoblue” tiene un formato en que se identifica la hora y la fecha en la que cada medición es realizada y esto se utiliza para diferenciarlos entre ellos. Ahora procederé a añadir las etiquetas para poder evaluar correctamente cada mes.

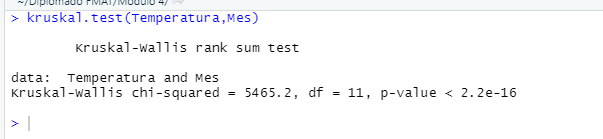


La primera columna donde esta el resaltado amarillo, fue añadida por mí para poder hacer el ANOVA correctamente, ya que necesita diferenciarse por los factores, que en este caso son los meses.

Ahora se tiene que checar si se cumplen los criterios para utilizar ANOVA, o en caso contrario utilizar Kruskal-Wallis. En las pruebas de normalidad en R Studio, no pude realizar el test de Shapiro-Wilk debido a que acepta como máximo 5000 datos, los cuales se quedan cortos con la cantidad de datos que manejamos (más de 8000 datos). El test de Liliefors muestra que la distribución no es normal (D = 0.017492, p-value = 2.356e-06 <0.05 ).





Debido a esto se procedió a utilizar Kruskal-Wallis, donde obtuvimos la siguiente tabla que indica una diferencia significativa entre los factores (Mes) con respecto a la variable de respuesta (Temperatura).

Al utilizar la prueba de Dunne obtuvimos los siguientes valores, los cuales muestran que hay diferencias significativas entre casi cada temperatura que hay por mes en Basel, Suiza. Estos fueron guardados en un Excel y se añadió una columna extra que decía si era significativo con un alfa de 0.05).

|  |  |  |  |
| --- | --- | --- | --- |
|  | Comparaciones | P-Valor | Significativo con α = 0.05 |
| 1 | ABRIL - AGOSTO | 1.35E-59 | SI |
| 2 | ABRIL - DICIEMBRE | 1.93E-139 | SI |
| 3 | AGOSTO - DICIEMBRE | 0 | SI |
| 4 | ABRIL - ENERO | 3.55E-139 | SI |
| 5 | AGOSTO - ENERO | 0 | SI |
| 6 | DICIEMBRE - ENERO | 1 | NO |
| 7 | ABRIL - FEBRERO | 6.51E-80 | SI |
| 8 | AGOSTO - FEBRERO | 3.11E-277 | SI |
| 9 | DICIEMBRE - FEBRERO | 6.86E-08 | SI |
| 10 | ENERO - FEBRERO | 5.54E-07 | SI |
| 11 | ABRIL - JULIO | 7.38E-56 | SI |
| 12 | AGOSTO - JULIO | 1 | NO |
| 13 | DICIEMBRE - JULIO | 0 | SI |
| 14 | ENERO - JULIO | 0 | SI |
| 15 | FEBRERO - JULIO | 3.48E-269 | SI |
| 16 | ABRIL - JUNIO | 1.76E-18 | SI |
| 17 | AGOSTO - JUNIO | 3.22E-11 | SI |
| 18 | DICIEMBRE - JUNIO | 2.32E-258 | SI |
| 19 | ENERO - JUNIO | 2.69E-260 | SI |
| 20 | FEBRERO - JUNIO | 2.11E-174 | SI |
| 21 | JULIO - JUNIO | 1.36E-09 | SI |
| 22 | ABRIL - MARZO | 8.72E-75 | SI |
| 23 | AGOSTO - MARZO | 6.90E-272 | SI |
| 24 | DICIEMBRE - MARZO | 6.13E-11 | SI |
| 25 | ENERO - MARZO | 6.39E-10 | SI |
| 26 | FEBRERO - MARZO | 1 | NO |
| 27 | JULIO - MARZO | 8.77E-264 | SI |
| 28 | JUNIO - MARZO | 1.21E-168 | SI |
| 29 | ABRIL - MAYO | 1 | NO |
| 30 | AGOSTO - MAYO | 5.88E-48 | SI |
| 31 | DICIEMBRE - MAYO | 1.95E-162 | SI |
| 32 | ENERO - MAYO | 1.35E-162 | SI |
| 33 | FEBRERO - MAYO | 4.29E-97 | SI |
| 34 | JULIO - MAYO | 1.40E-44 | SI |
| 35 | JUNIO - MAYO | 4.78E-12 | SI |
| 36 | MARZO - MAYO | 1.04E-91 | SI |
| 37 | ABRIL - NOVIEMBRE | 2.63E-53 | SI |
| 38 | AGOSTO - NOVIEMBRE | 4.48E-227 | SI |
| 39 | DICIEMBRE - NOVIEMBRE | 7.64E-21 | SI |
| 40 | ENERO - NOVIEMBRE | 1.21E-19 | SI |
| 41 | FEBRERO - NOVIEMBRE | 0.017798795 | SI |
| 42 | JULIO - NOVIEMBRE | 9.82E-220 | SI |
| 43 | JUNIO - NOVIEMBRE | 1.03E-134 | SI |
| 44 | MARZO - NOVIEMBRE | 0.383218191 | NO |
| 45 | MAYO - NOVIEMBRE | 1.98E-67 | SI |
| 46 | ABRIL - OCTUBRE | 1.17E-10 | SI |
| 47 | AGOSTO - OCTUBRE | 3.42E-123 | SI |
| 48 | DICIEMBRE - OCTUBRE | 1.04E-74 | SI |
| 49 | ENERO - OCTUBRE | 1.29E-73 | SI |
| 50 | FEBRERO - OCTUBRE | 5.22E-33 | SI |
| 51 | JULIO - OCTUBRE | 9.36E-118 | SI |
| 52 | JUNIO - OCTUBRE | 2.51E-58 | SI |
| 53 | MARZO - OCTUBRE | 3.99E-29 | SI |
| 54 | MAYO - OCTUBRE | 2.39E-17 | SI |
| 55 | NOVIEMBRE - OCTUBRE | 1.94E-16 | SI |
| 56 | ABRIL - SEPTIEMBRE | 1.12E-19 | SI |
| 57 | AGOSTO - SEPTIEMBRE | 2.66E-10 | SI |
| 58 | DICIEMBRE - SEPTIEMBRE | 1.10E-262 | SI |
| 59 | ENERO - SEPTIEMBRE | 1.04E-264 | SI |
| 60 | FEBRERO - SEPTIEMBRE | 5.87E-178 | SI |
| 61 | JULIO - SEPTIEMBRE | 9.65E-09 | SI |
| 62 | JUNIO - SEPTIEMBRE | 1 | NO |
| 63 | MARZO - SEPTIEMBRE | 3.37E-172 | SI |
| 64 | MAYO - SEPTIEMBRE | 4.95E-13 | SI |
| 65 | NOVIEMBRE - SEPTIEMBRE | 7.19E-138 | SI |
| 66 | OCTUBRE - SEPTIEMBRE | 1.98E-60 | SI |

Estos datos muestran que los únicos meses que no tuvieron diferencias significativas de temperatura fueron ~~j~~unio y septiembre, marzo y noviembre, abril y mayo, febrero y marzo, agosto y julio, y diciembre y enero. Todos los demás meses presentaron diferencias significativas. Las temperaturas medias de cada mes ordenadas de menor a mayor son las siguientes (los meses con la misma letra en la columna de grupo se considera que tienen temperaturas iguales).

Mes Temperatura Grupo

DICIEMBRE 5.331043 a

ENERO 5.517440 a

FEBRERO 7.686957 b

MARZO 7.943015 bc

NOVIEMBRE 8.949010 c

OCTUBRE 12.057494 d

ABRIL 14.536855 e

MAYO 15.291558 e

JUNIO 18.109189 f

SEPTIEMBRE 18.462518 f

JULIO 21.128604 g

AGOSTO 21.714259 g

Ya teniendo esta información podemos decir que los meses ordenados de menor a mayor temperatura significativa son:

1.- Diciembre y Enero

2.- Febrero y Marzo

3.- Noviembre

4.- Octubre

5.- Abril y Mayo

6.- Junio y Septiembre

7.- Julio y Agosto

**Conclusiones**

Con esto podemos concluir que los meses de diciembre a marzo son los mejores para personas que les gustan las actividades a bajas temperaturas, mientras que las actividades en climas más cálidos se pueden realizar de mejor manera desde junio hasta septiembre. Esto nos servirá para poder informar a los turistas que meses serían los mejores para visitar Basilea dependiendo de las actividades que desean realizar, y en caso que no puedan ajustar su tiempo vacacional, saber que tipo de actividades tienen disponibles para hacer.

**Bibliografía**

A student’s guide to Analysis of Variance,Autores: Maxwell J. Roberts y Riccardo Russo

Estadística elemental 11a edición Johnson Kuby

A student’s guide to Analysis of Variance,Autores: Maxwell J. Roberts y Riccardo Russo

An introduction to non parametrics statistics, Autor: John E. Kolassa

1964. Multiple comparisons using rank sums. Technometrics

<https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/anova/supporting-topics/basics/what-is-anova/>

<https://www.statistics.com/glossary/post-hoc-tests/>

<https://stackoverflow.com/questions/14359726/sort-matrix-according-to-first-column-in-r>

<https://stackoverflow.com/questions/23395366/mean-by-factor-by-level>

<http://www.r-tutor.com/r-introduction/vector/combining-vectors>

<https://www.myswitzerland.com/es/destinos/basilea/>

<https://www.oyster.com/es/articles/best-time-to-visit-switzerland/>

**Anexo**

**ANOVA**

La técnica de análisis de varianza (ANOVA) se usará para poner a prueba una hipótesis nula acerca de varias medias, por ejemplo, .

Al usar la técnica anterior para hipótesis en torno a dos medias, podrías poner a prueba varias hipótesis si cada una enunciara una comparación de dos medias. Por ejemplo, podrías poner a prueba

;

Para poner a prueba la hipótesis nula, , de que las cinco medias son iguales, tendrías que poner a prueba cada una de estas 10 hipótesis usando la técnica anterior. El rechazo de alguna de las 10 hipótesis en torno a dos medias causaría el rechazo de la hipótesis nula de que las cinco medias son iguales. Si se fallara en rechazar las 10 hipótesis, se fallaría en rechazar la hipótesis nula principal. Al poner a prueba de esta forma, la tasa global de error tipo I se volvería mucho más grande que el valor de α asociado con una sola prueba. Por tanto, las técnicas ANOVA permiten poner a prueba la hipótesis nula (todas las medias son iguales) contra la hipótesis alternativa (al menos un valor de media es diferente) con un valor específico de α.

Los experimentos ANOVA pueden ser muy complejos, dependiendo de la situación. La discusión se restringirá al diseño experimental más básico: la ANOVA de un solo factor.

Para ejecutar un ANOVA, debe tener una variable de respuesta continua y al menos un factor categórico con dos o más niveles. Los análisis ANOVA requieren datos de poblaciones que sigan una distribución aproximadamente normal con varianzas iguales entre los niveles de factores. Sin embargo, los procedimientos de ANOVA funcionan bastante bien incluso cuando se viola el supuesto de normalidad, a menos que una o más de las distribuciones sean muy asimétricas o si las varianzas son bastante diferentes. Las transformaciones del conjunto de datos original pueden corregir estas violaciones.

**Kruskal-Wallis**

Se puede construir un caso simple del estadístico de rango multivariado general eligiendo los puntajes para los estadísticos de rango como la identidad, con los rangos mismos como puntajes. Kruskal y Wallis (1952) introdujeron la prueba que rechaza la hipótesis nula de Man-Whitney-Wilcoxon de distribuciones iguales. cuando el estadístico de prueba excede el cuantil apropiado del valor nulo de la distribución . Aplican esto con puntajes iguales a los rangos. Utilizando:



La estadística se simplifica a exceder el valor crítico apropiado.

Aquí, nuevamente, es el rango de dentro de la muestra combinada, y .

Utilizamos esta prueba no paramétrica en lugar del ANOVA cuando no se cumplen las condiciones de igualdad de varianzas o el supuesto de normalidad. En esta prueba se checa si las k muestras provienen de poblaciones idénticas con la misma mediana.

**El problema de múltiples comparaciones**

La razón para no realizar múltiples pruebas-t o de algún otro tipo para comparar cada promedio es que múltiples comparaciones estadísticas incrementan la probabilidad de hacer un Error de Tipo 1. El Error de Tipo 1 ocurre cuando la hipótesis nula es rechazada erróneamente (es verdadera) mientras que si la hipótesis nula es falsa y esta no se rechaza, se le conoce como Error de Tipo 2. Dependiendo del tipo de estudio estos errores tienen una mayor importancia. Mientras que en casi todas las áreas el error de tipo 1 es más importante que el de tipo 2 (ya que se desprende una conclusión al rechazar la hipótesis nula), hay excepciones como la medicina, ya que si por ejemplo hay un error de tipo 2 y se le comunica a una persona que no tiene una enfermedad cuando realmente si la tiene, es peligroso.

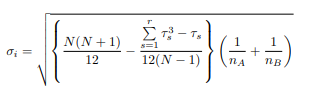
La razón de que se incremente el error de tipo 1 en múltiples comparaciones estadísticas (tasa de error por familia) es que el error puede aparecer en muchos más lugares, al igual que múltiples veces. Esto necesita ser tomado en cuenta con los tests post-hoc y las comparaciones planeadas.

Se pueden utilizar comparaciones planeadas en lugar de los tests de tipo post-hoc si solo se comparan unas medias específicas en lugar de todas. Esto se hace debido a que las comparaciones planeadas son más potentes que los tests post-hoc.

**Test de Dunne**

Es un método no paramétrico de comparación múltiple. El estadístico de la prueba z de Dunn se aproxima a los estadísticos exactos de la prueba de suma de rangos mediante el uso de los rangos medios del resultado en cada grupo de la prueba anterior de Kruskal-Wallis (, donde es la suma de rangos y es la muestra tamaño para el i-ésimo grupo) y basar la inferencia en las diferencias en los rangos medios en cada grupo. Para comparar el grupo A con el grupo B, calculamos

where i is one of the 1 to m multiple comparisons, , y es la desviación estándar de , dada por



donde N es el número total de observaciones en todos los grupos, r es el número de rangos empatados y es el número de observaciones empatadas en el s-ésimo valor vinculado específico. Cuando no hay empates, el término con la suma en el denominador es igual a cero, y el cálculo de (2) simplifica considerablemente.