Nearest Neighbor Statistic in R

Venn Datagram

2024-07-24

Nearest Neighbor Statistic Z-Score

$$z = \frac{\bar{r}_0 - r_e}{s_e}$$

Mean observed distance

$$\bar{r}_0 = \frac{\sum_{i=1}^{n} d_i}{n}$$

```
library(geosphere)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df <- read.csv(
  "Public_School_Characteristics_-_Current.csv",
  encoding = "UTF-8")
```

```
#Let's Only Grab the Locations This Time
t.df <- df %>% filter(
  STABR == "VA",
  is.na(TOTAL) == FALSE,
  SCHOOL_LEVEL=="High") %>% select(
    LATCOD,LONCOD)
head(t.df)
```

```
##     LATCOD    LONCOD
## 1 38.15066 -79.06397
## 2 37.94280 -75.36380
## 3 37.67920 -75.72560
## 4 37.92530 -75.55010
## 5 38.07620 -78.50120
## 6 38.04071 -78.48276
```

```
find.nearest.neighbor <- function(lon, lat, df) {
  distances <- distm( #Calculates pairwise distance between multiple points
    cbind(lon, lat), #YOU MUST DO LON LAT FOR HAVERSINE NOT LAT LON
```

```
    cbind(df$LONCOD, df$LATCOD),
    fun = distHaversine) #haversine distance like last time since lat/lon data

  distances[distances == 0] <- NA  # Exclude the point itself

  nearest.neighbor <- which.min(distances) #index of nearest neighbor

  return(nearest.neighbor)
}
```

```
#Get ID of nearest neighbor point, then add it to row in data
t.df <- t.df %>%
  rowwise() %>%
  mutate(
    n.neighbor = find.nearest.neighbor(LONCOD,LATCOD, t.df),
    n.neighbor.coord.lon = t.df$LONCOD[n.neighbor],
    n.neighbor.coord.lat = t.df$LATCOD[n.neighbor]
  ) %>%
  ungroup()
```

```
# Calculate the distance between each point and its nearest neighbor using sapply
distances <- sapply(1:nrow(t.df), function(i) {
  distHaversine(
    c(t.df$LONCOD[i], t.df$LATCOD[i]),
    c(t.df$n.neighbor.coord.lon[i], t.df$n.neighbor.coord.lat[i])
  )
})

# Add the distances to the dataframe
t.df$distance_to_nearest_neighbor <- distances
```

```
r.bar.0 <- sum(t.df$distance_to_nearest_neighbor) / dim(t.df)[1]
r.bar.0 # 9458.271 meters
```

```
## [1] 9458.271
```

Expected Value for Distance Between Points

$$r_e = \frac{1}{2\sqrt{\frac{n}{A}}} = \frac{1}{2\sqrt{\lambda}}$$

n = number of points in the study area

A = study area

Derivation Suppose we are looking at the distance between points and we have N total points in a study area of A. We could say the rate of occurrence for points is $\lambda = \frac{n}{A}$ for our study. We can treat this as the number of events occurring on an interval, which sounds like Poisson. We assume that points are independent and occur at a constant distance in our fixed area.

First, we rule out seeing 0 points in our area. So if we treat μ as the rate of number of points in our area, we can substitute mu for $A\lambda$ since we expect n points in our area. If you are confused about that, think like this: $\lambda = \frac{n}{A}$ and $A\lambda = A\frac{n}{A} = n$.

$f(x) = \frac{e^{-\mu}\mu^x}{x!}$ so when x = 0,

$f(0) = \frac{e^{-(A\lambda)}(A\lambda)^0}{0!} = \frac{e^{-(A\lambda)}*1}{1} = e^{-A\lambda}$

Poisson is a discrete distribution, so we can use that to find the probability of a point existing in the area:

$Pr(X > 0) = 1 - Pr(X = 0) = 1 - f(0) = 1 - e^{-A\lambda}$

Let's think of area as a circle. We are looking for points within our area of πr^2.

The proportion of distances to our nearest neighbor $\leq r$ is:

$F(r) = 1 - e^{-(\pi r^2)\lambda}$

Next, we get the probability distribution of r, which is the derivative of the above with respect to r:

$$f(r) = \frac{d}{dr}F(r) = 2\lambda\pi r e^{-(\pi r^2)\lambda}$$

We know the expected value of a distribution is

$$\int_{-\infty}^{\infty} xf(x)dx$$

so to find the expected value r_e we can substitute this in:

$$\int_0^{\infty} rf(r)dr = \int_0^{\infty} r(2\lambda\pi r e^{-(\pi r^2)\lambda})dr = \int_0^{\infty} 2\lambda\pi r^2 e^{-(\pi r^2)\lambda}dr$$

We can then perform u-substitution, where $u = \lambda\pi r^2$ and $du = (2\lambda\pi r)dr$. This also means $r = \sqrt{\frac{u}{\lambda\pi}}$ and $dr = \frac{du}{(2\lambda\pi r)} = \frac{du}{(2\lambda\pi\sqrt{\frac{u}{\lambda\pi}})} = \frac{du}{2\lambda\pi\sqrt{\frac{1}{\lambda\pi}}\sqrt{u}} = \frac{du}{2\lambda\pi\frac{1}{\sqrt{\lambda\pi}}\sqrt{u}} = \frac{du}{2(\lambda\pi)^1(\lambda\pi)^{-\frac{1}{2}}\sqrt{u}} = \frac{du}{2(\lambda\pi)^{\frac{1}{2}}\sqrt{u}} = \frac{du}{2\sqrt{\lambda\pi}\sqrt{u}}$. This also means $r^2 = \frac{u}{\lambda\pi}$ so:

$$\int_0^{\infty} 2\lambda\pi(r^2)e^{-(\lambda\pi r^2)}dr = \int_0^{\infty} 2\lambda\pi(\frac{u}{\lambda\pi})e^{-(u)}(\frac{du}{2\sqrt{\lambda\pi}\sqrt{u}}) = \int_0^{\infty} \frac{1}{\sqrt{\lambda\pi}} * \frac{u}{\sqrt{u}}e^{-u}du$$

We can simplify this even further:

$$= \frac{1}{\sqrt{\lambda\pi}}\int_0^{\infty} u^1 u^{\frac{-1}{2}}e^{-u}du = \frac{1}{\sqrt{\lambda\pi}}\int_0^{\infty} u^{\frac{1}{2}}e^{-u}du$$

If we are looking at this, the second part looks like the Gamma Function where $x = \frac{3}{2}$:

$$\Gamma(x) = \int_0^{\infty} y^{x-1}e^{-y}dy$$

There are two relevant properties of the gamma function:

1. $\Gamma(z+1) = z\Gamma(z)$
2. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

Rabbit Hole Derivation for Point 2 (If you are interested)

There are two things I love: going off on random tangents and making connections between interesting topics, so if you are as passionate about mathematics and surprise relationships between functions and want a break from the Nearest Neighbors Statistic, continue reading. Else, feel free to skip this part.

$$\Gamma(\frac{1}{2}) = \int_0^{\infty} y^{\frac{1}{2}-1}e^{-y}dy = \int_0^{\infty} y^{\frac{-1}{2}}e^{-y}dy$$

We can use u-substitution where $u = y^{\frac{1}{2}}$ so $y = u^2$, and $du = \frac{1}{2}y^{-\frac{1}{2}}dy$ so $2du = y^{-\frac{1}{2}}dy$ so:

$$\int_0^\infty 2e^{-u^2}du$$

Then by the Rule for Definite Integrals of Even Functions:

$$\int_{-\infty}^\infty e^{-u^2}du$$

Let's take a moment and think for a second. This is the Gaussian function $f(x) = e^{-x^2}$ inside the integral.

You are likely familiar with the parametric extension of the Gaussian for a Normal Distribution (if this sentence sounded like gibberish to you, click the link I attached and you'll get it). Here is the Normal Distribution function:

$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ or rearranged as $\frac{1}{\sigma\sqrt{2\pi}}e^{-(\frac{x-\mu}{\sigma\sqrt{2}})^2}$.

What is so special about the area under the curve? It's 1! If we were to think about this as an integral, it would look like:

$1 = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}}e^{-(\frac{x-\mu}{\sigma\sqrt{2}})^2}dx$

and with u-substitution where $u = \frac{x-\mu}{\sigma\sqrt{2}}$ and $du = \frac{1}{\sigma\sqrt{2}}dx$

$1 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}}e^{-u^2}du = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-u^2}du$

Isn't that interesting? The area under the Normal Distribution curve is 1, but it looks like it can also be calculated by dividing the integral of the Gaussian Function by $\sqrt{\pi}$. That means the integral of the Gaussian Function divided by $\sqrt{\pi}$ is 1, so the integral of the Gaussian Function equals $\sqrt{\pi}$ itself. The integral of the Gaussian function also equals $\Gamma(\frac{1}{2})$, meaning $\Gamma(\frac{1}{2}) = \sqrt{\pi}$!

Return to Main Topic

Coming back, we have our Gamma Function where $x = \frac{3}{2}$:

$$\Gamma(x) = \int_0^\infty y^{x-1}e^{-y}dy$$

and two relevant properties:

1. $\Gamma(z+1) = z\Gamma(z)$
2. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

So:

$\Gamma(\frac{3}{2}) = \Gamma(\frac{1}{2}+1) = \frac{1}{2}\Gamma(\frac{1}{2}) = \frac{1}{2}\sqrt{\pi}$

To put this all together, this was our previous final result:

$$\frac{1}{\sqrt{\lambda\pi}}\int_0^\infty u^{\frac{1}{2}}e^{-u}du$$

But now we can simplify this to:

$$\frac{1}{\sqrt{\lambda\pi}} * \Gamma(\frac{3}{2}) = \frac{1}{\sqrt{\lambda\pi}} * (\frac{1}{2}\sqrt{\pi}) = \frac{1}{2\sqrt{\lambda}}$$

where λ is the rate of occurrence of points in our area.

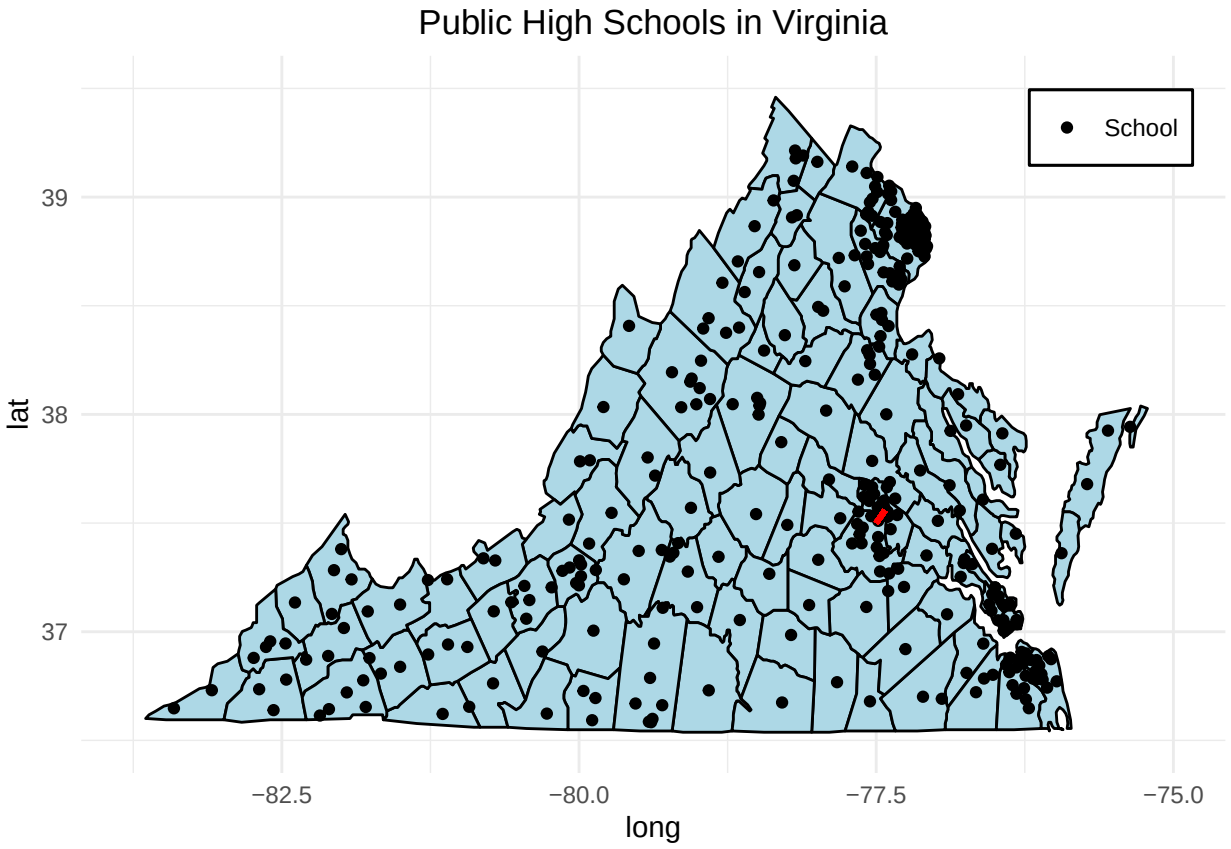
Let's do this in R!

First, we need a λ so we need our n and A. Virginia is massive and we do not have many points, so the denominator is going to be very small, making our fraction very big. That makes sense, since the expected distance between points should be large (we are using meters for all measurements because that is the unit for the *distHaversine()* function.

```
VA.in.meters.2 <- 110785670000

n <- nrow(t.df)

lambda <- n / VA.in.meters.2

r.e <- 1/(2*sqrt(lambda))

r.e #[1] 9217.285 meters
```

```
## [1] 9217.285
```

It looks like the expected distance between schools is around 9.2 km or 5.27 miles. Now, before we continue any further, lets look at the area around Richmond Virginia and draw a line to show 5.27 miles and see how realistic this is.



Hmmm. We are expecting all the points to be the distance of the red line from each other. I'm **"SURE"** we won't find any clusters (air quotes).

Standard Error for Expected Distance Between Points

$$s_e = \frac{0.261}{\sqrt{\frac{n^2}{A}}}$$

This was what I saw when I read the formula for the first time. My immediate thought was "there's no way the value 0.261 is not an approximation of some other value. Where does it come from?" I was right, it's merely an approximation. I'll show you its origin story.

Derivation We are looking for the standard error for expected distance, so this formula is going to come into play:

$V(r_e) = E(r_e^2) - [E(r_e)]^2$

We already know $E(r_e)$ from the previous exercise, which is $\frac{1}{2\sqrt{\lambda}}$. We calculated the first moment with $E(r) = \int_0^\infty rf(r)dr$ so we can calculate the second moment with $E(r^2) = \int_0^\infty r^2 f(r)dr$.

$E(r_e^2) = \int_0^\infty r^2 f(r)dr = \int_0^\infty r^2(2\lambda\pi re^{-\pi r^2\lambda})dr$

The $f(r)$ comes from our previous equation's derivation.

If we perform the same u-substitution as before ($u = \lambda\pi r^2$ and $du = (2\lambda\pi r)dr$), we can substitute $r^2 = \frac{u}{\lambda\pi}$ and $\frac{du}{2\lambda\pi} = rdr$

$$E(r_e^2) = \int_0^\infty 2\lambda\pi(r^2)e^{-(\lambda\pi r^2)}(rdr) = \int_0^\infty 2\lambda\pi(\frac{u}{\lambda\pi})e^{-(u)}(\frac{du}{2\lambda\pi}) = \frac{1}{\lambda\pi}\int_0^\infty ue^{-u}du$$

And again, we can use the Gamma Function, but this time, we can use it's other property $\Gamma(z) = (z-1)!$.

$$E(r_e^2) = \frac{1}{\lambda\pi}\Gamma(2) = \frac{1}{\lambda\pi}(2-1)! = \frac{1}{\lambda\pi}(1)! = \frac{1}{\lambda\pi}$$

If we substitute this into our equation for $V(r_e)$ we get

$V(r_e) = E(r_e^2) - [E(r_e)]^2 = \frac{1}{\lambda\pi} - [\frac{1}{2\sqrt{\lambda}}]^2$

This simplifies to

$\frac{1}{\lambda\pi} - [\frac{1}{4\lambda}] = \frac{4}{4\lambda\pi} - [\frac{\pi}{4\lambda\pi}] = \frac{4-\pi}{4\lambda\pi} = \frac{4-\pi}{4\pi} * \frac{1}{\lambda}$

Why did I write it like this? I will show you in a second.

We calculated the variance of r_e. In order to get the standard deviation, we need to square root the value:

$\sqrt{\frac{4-\pi}{4\pi} * \frac{1}{\lambda}} = \sqrt{\frac{4-\pi}{4\pi}} * \sqrt{\frac{1}{\lambda}}$

What is this... approximately?

$\sqrt{\frac{4-\pi}{4\pi}} * \sqrt{\frac{1}{\lambda}} \approx 0.261 * \sqrt{\frac{1}{\lambda}} = \frac{0.261}{\sqrt{\lambda}} = \frac{0.261}{\sqrt{\frac{n}{A}}}$

We did it!... or did we? The n is supposed to be squared. Why does the equation above yield this answer and not n^2 like I showed in the defining equation?

We calculated the standard deviation rather than the standard error above, and we need to calculate the standard error instead. There is a difference between the two.

In order to show you what I mean, let's talk about properties of Variance.

There are two that are relevant:

1. $Var(X+Y) = Var(X) + Var(Y) + Cov(X,Y)$

2. $Var(aX) = a^2 Var(X)$

where a is a constant.

What we calculated before was variance of r_e and standard deviation by taking the root. In reality, we are not calculating r_e but $\bar{r}_e$. This is our sample mean, which results in a slightly different value. Here is what I mean:

$Var(\bar{x}) = Var(\frac{\sum x}{n}) = \frac{1}{n^2} Var(\sum x) = \frac{1}{n^2} \sum Var(x)$

Since our problem has points that we consider identically distributed, $Cov(X, Y)$ is going to be 0, so $Var(X + Y) = Var(X) + Var(Y)$. Thus, we get $\frac{n}{n^2} Var(x)$, which is just $\frac{1}{n}\sigma^2$. If we square root it, then we get $\frac{\sigma}{\sqrt{n}}$, which is the sample error.

Notice that it multiplies σ by $\frac{1}{\sqrt{n}}$. This means our standard error is going to be the standard deviation multiplied by the above fraction. We calculated the standard deviation as:

$\approx \frac{0.261}{\sqrt{\frac{n}{A}}}$

so we multiply by $\frac{1}{\sqrt{n}}$ and get the final standard error: $\frac{0.261}{\sqrt{\frac{n^2}{A}}}$.

In R, this is very simple:

```
VA.in.meters.2 <- 110785670000

n <- nrow(t.df)

lambda.n <- n^2 / VA.in.meters.2

s.e <- sqrt((4-pi)/(4*pi)) * (1/(sqrt(lambda.n)) )

s.e #[1] 266.8492 meters
```

```
## [1] 266.8492
```

Put Everything Together

I have one thing to say (sing): I can z clearly now, the time has come! If you did not understand the song reference, it's alright, we can simply do the calculation.

We have this formula:

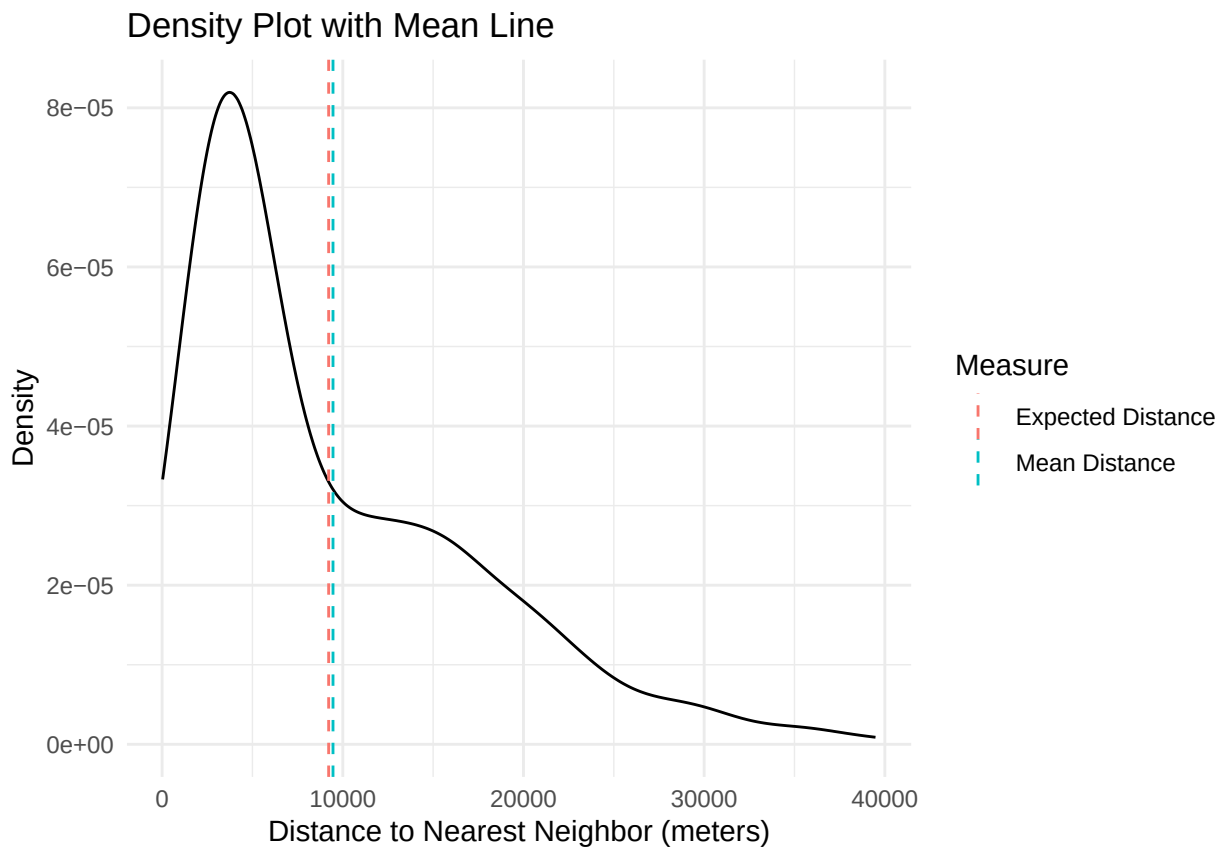
$$z = \frac{\bar{r}_0 - r_e}{s_e}$$

Let's see the z-score that we get when we put in our calculated values:

```
z <- (r.bar.0 - r.e) / s.e

z #[1] 0.9030797
```

```
## [1] 0.9030797
```

Well, that was disappointing. The z is so small. It does not come close to disproving a hypothesis of no clusters. The alpha that we set would have to be pretty big and nowhere close to 0.05 (not even 0.1). Let's plot the distances of all the points from their nearest neighbors to understand what happened:



It seems like the most successful study is going to be the one where the peak density is even greater than the other areas, so we would need to see an even more increased skew with other points further away from the clusters we visualized. Other than that, this is a real lesson in the boundary selection: the area we select is going to determine which points are nearest, since we might close off a point that was the nearest neighbor to the one in our study region, leading us to a different conclusion. Anyways, it's alright. We will talk about other strategies that can be more successful, but this was a great learning experience!