Descriptive Statistics

Measurements of *location* and *spread* of data

BDSiC – Day 3 A

Agenda:

- Mean, mode, median
- Variability, variation, range
- Accuracy/Bias and Precision/Spread
- Data types and their common visualizations: scatterplots, histograms, boxplots/violin plots, bar plots (mosaic plots)
- Interpretation of popular plots in genomics

You are considering buying a house in a certain neighbourhood. You find a potential house and, to appeal to perceived snobbiness as you are making your decision, your realtor mentions that the **average income in this neighbourhood is $100,000 per year.**

You buy the house.

A year later, the same realtor knocks on your door, this time acting as a representative of the neighbourhood taxpayers' association. He would like you to sign a petition to decrease property taxes because, he says, the residents can't afford an increase in property taxes since the **average family income in the neighbourhood is only $25,000 per year.**

How is this possible, if the realtor is telling the truth, and no one in the neighbourhood has moved or changed jobs in the last year?

The two common descriptions of data:

1. **Location:**
   - Central Tendency
   - Where is the weight of the data?

Average

2. **Spread:**
   - How far apart are the data points? Especially: how far apart are the largest and smallest data points?

Range

You will also see:

1. **Skew –** The third standardized moment; positive or negative skew. The shape of the distribution is not symmetric.
2. **Kurtosis** – The fourth standardized moment; sort of 'peakness' of the distribution (fatness of the tails)

A story about central location of the data

Waiter	$35,000
Cook	$30,000
Dishwasher	$25,000
Customer 1	$80,000
Customer 2	$50,000
Customer 3	$30,000
Customer 4	$45,000

"Average" is approx. **$42,143**

"Average" is $**125,000,037**

Waiter	$35,000
Cook	$30,000
Dishwasher	$25,000
Customer 1	$80,000
Customer 2	$50,000
Customer 3	$30,000
Customer 4	$45,000
Software or Social Engineer	$1,000,000,000

$35,000
$30,000
$25,000
$80,000
$50,000
$30,000
$45,000

Reorder data →

$25,000
$30,000
$30,000
$35,000
$45,000
$50,000
$80,000

(Arithmetic) **Mean** = $\frac{\sum_1^n x_i}{n}$

Median = middle value (odd), mean of middle value (even)

Mode = most frequent value

Concern: Skewness: in the generic distribution

$35,000
$30,000
$25,000
$80,000
$50,000
$30,000
$45,000
$1,000,000,000

Reorder data →

$25,000
$30,000
$30,000
$35,000
$45,000
$50,000
$80,000
$1,000,000,000

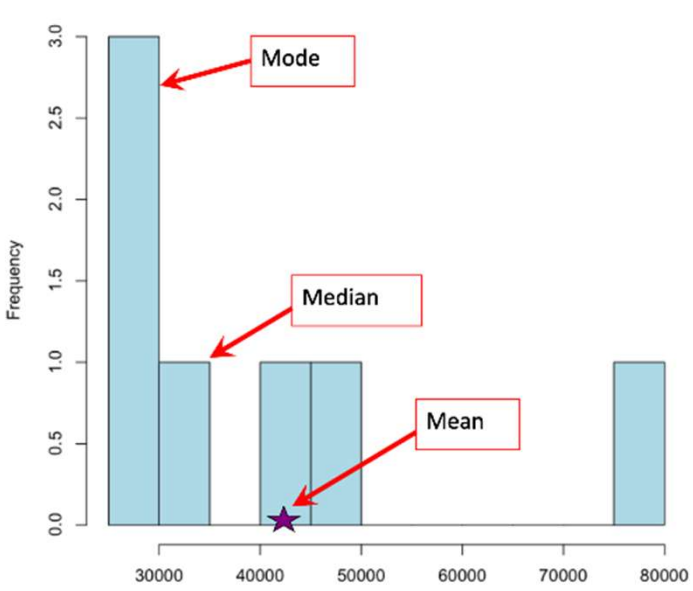
robustness

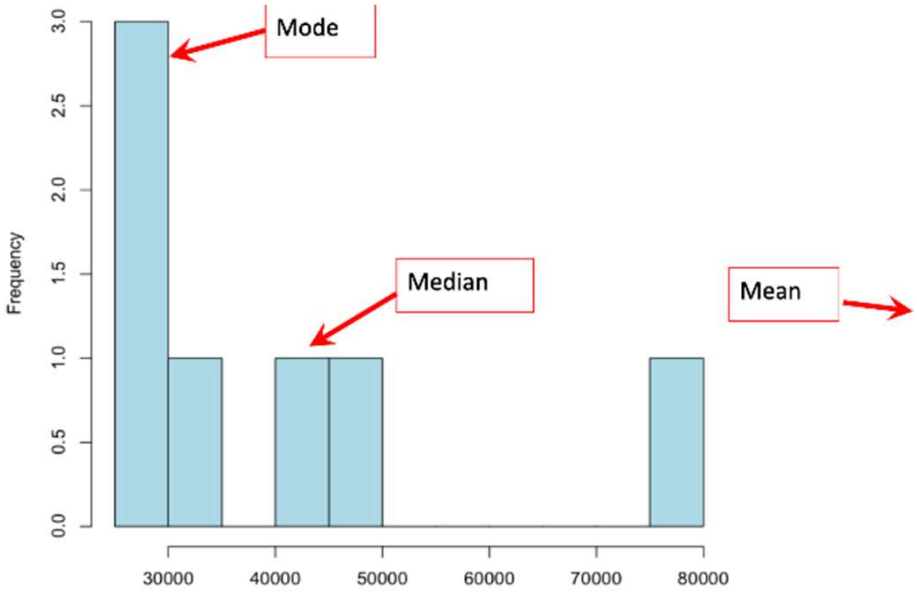
	Scenario 1	Scenario 2
mean	$42 143	$125,000,037
median	$35,000	$40,000
mode	$30,000	$30,000

Mean, Mode, and Median can give you different information and they have different benefits

- If the data are skewed or have an outlier, median is often a fairer reflection of the data
- Median can give quick information abut the data without having to calculate anything
- (arithmetic) mean can be a theoretical abstract (2.2 children per woman doesn't actually exist), but it allows you to use normal distribution to answer questions about the whole population

Track this





Will Rogers Phenomenon

"When the Okies left Oklahoma and moved to California, they raised the average intelligence level in both states."

https://en.wikipedia.org/wiki/Will_Rogers_phenomenon

Actual medical phenomenon: "medical stage migration"

Example: There were more COVID deaths among the vaccinated than the unvaccinated. In September 2022, 12,593 COVID deaths occurred in the United States. Of those, 39% were unvaccinated, while 61% were vaccinated. WHY?

Simple

Random Variables:

- Characteristics measured on individuals drawn from the population
- Value is not constant; it is subject to **VARIATION**
- **Categorical (Nominal, Ordinal)** or **Numeric (Discrete, Continuous)**

can be discretized

Types of data:

Categorical Variable

- AKA Class variables or Nominal variables
- They do not have magnitude on a numerical scale
- **Nominal**
  - Lack inherent order
- **Ordinal**
  - Inherent order **i.e. age (0-18, 19-30, 30-45, etc)**
- Ex: blood type, genotype, sex, state, survival (live or die), drug treatment (aspirin vs ibuprofen)

Quantitative Variables

- AKA Numerical variables
- Random Variable is a Quantitative variable
- **Continuous**
  - Ability to take any value ex.. Human weight, **age**
  - **They can be measured**
- **Discrete**
  - Spaces between possible values ex. Number of offspring, **age**
  - **They can be counted**

A research team is studying the health and fitness habits of a group of individuals. They collect the following data for each participant:

1. **Resting heart rate (beats per minute)**
2. **Favorite type of exercise (running, swimming, cycling, pilates, etc.)**
3. **Number of hours exercised per week**
4. **Body Mass Index (BMI)**
5. **Member status at a gym (yes or no)**

Which of the following (A, B, C, or D) correct classifies these variables:

A. Resting heart rate: **Nominal**
Favorite exercise: **Ordinal**
Number of hours of exercise per week: **Discrete**
BMI: **Continuous**
Membership status: **Nominal**

B. Resting heart rate: **Continuous**
Favorite exercise: **Nominal**
Number of hours of exercise per week: **Continuous**
BMI: **Continuous**
Membership status: **Categorical**

C. Resting heart rate: **Ordinal**
Favorite exercise: **Nominal**
Number of hours of exercise per week: **Continuous**
BMI: **Ordinal**
Membership status: **Nominal**

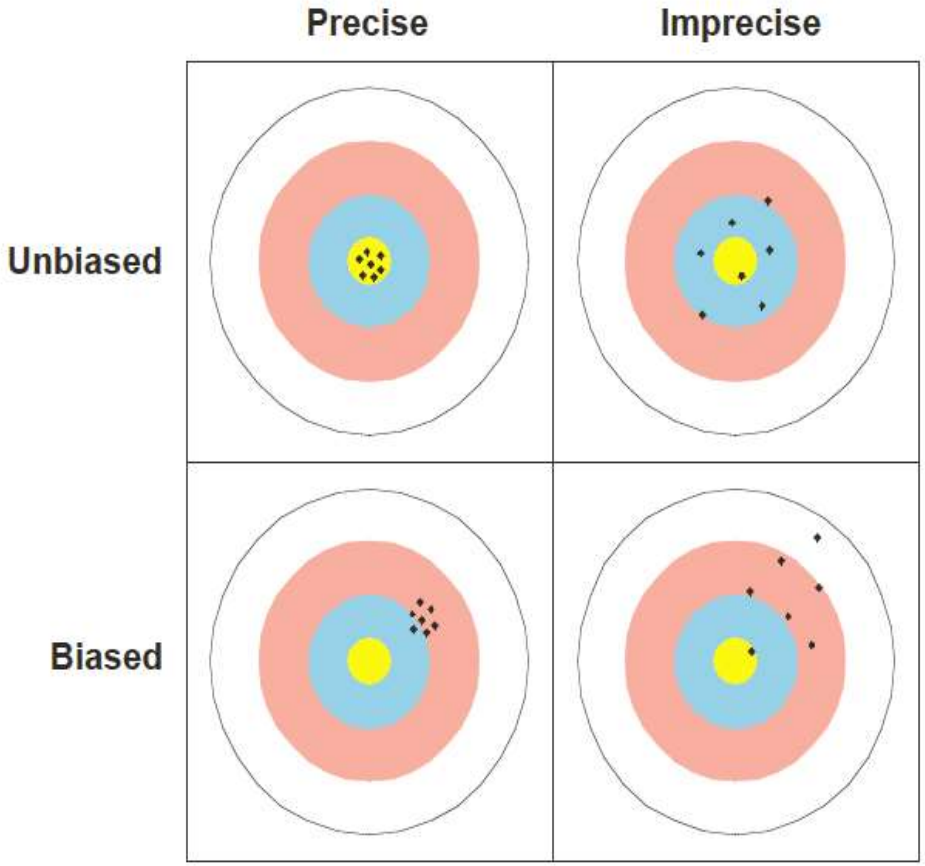
D. Resting heart rate: **Discrete**
Favorite exercise: **Continuous**
Number of hours of exercise per week: **Discrete**
BMI: **Continuous**
Membership status: **Ordinal**

Populations
have
PARAMETERS

- Represented by Greek Letters
- **μ; σ**

Samples
have
E***S***TIMATES

- Represented by Roman Letters
- $\bar{x}$ **; s**



Two major considerations:

1. Accuracy/biased

Bias:

a systematic discrepancy between estimates and the true population characteristic

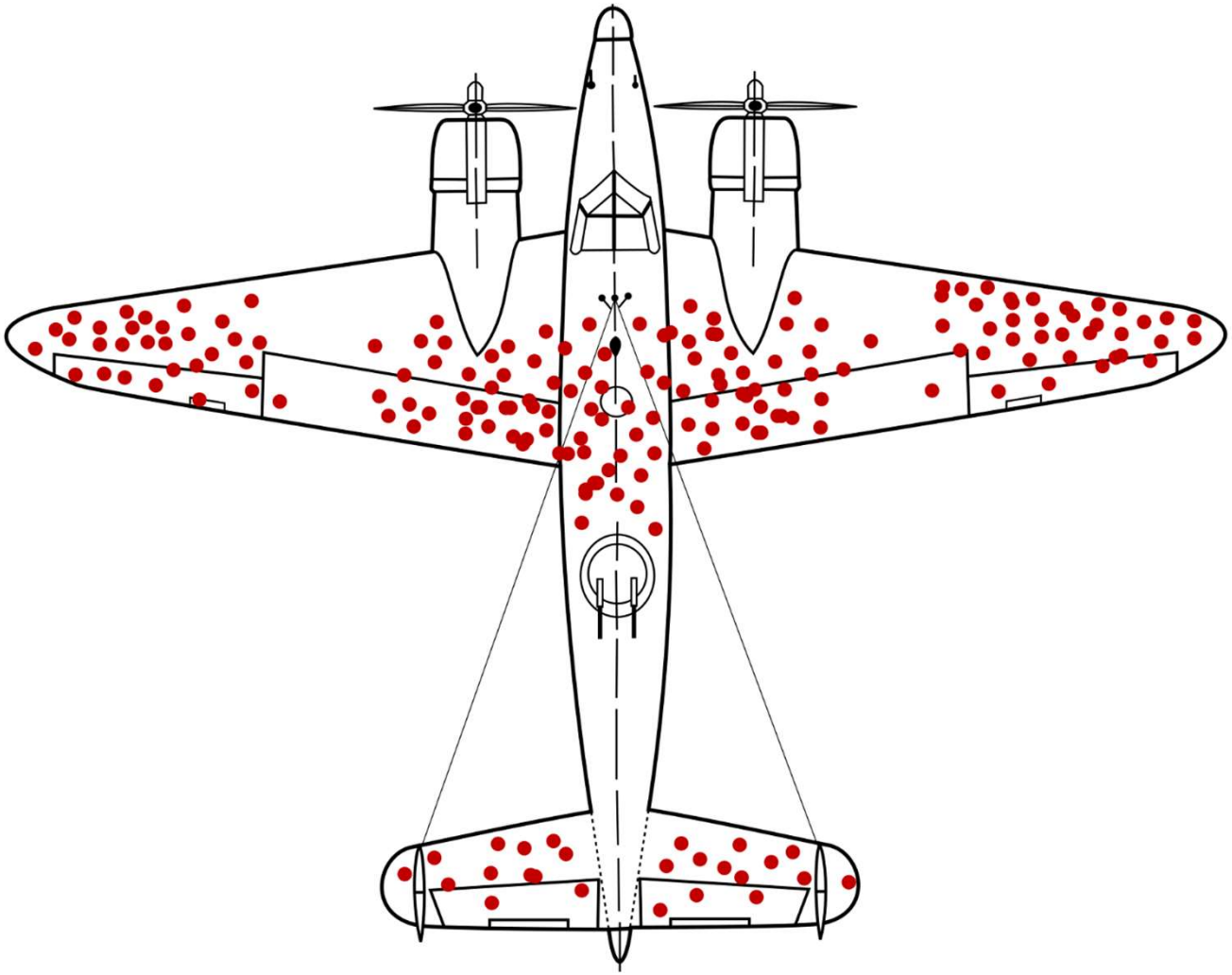
2. Precision/Spread

- Low Sampling Error, high precision

$$SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

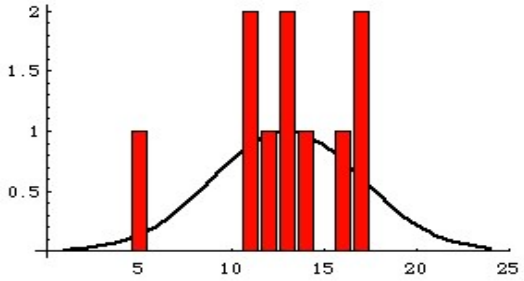
To address these, you typically need:

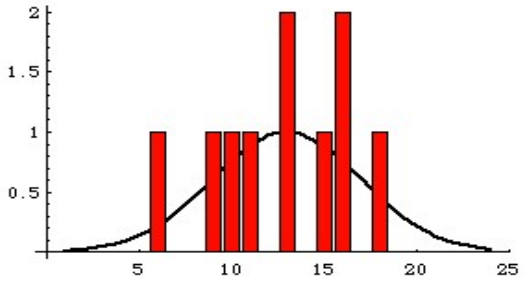
1. A sufficiently large sample
2. Randomly Sampled data points that are independent of each other

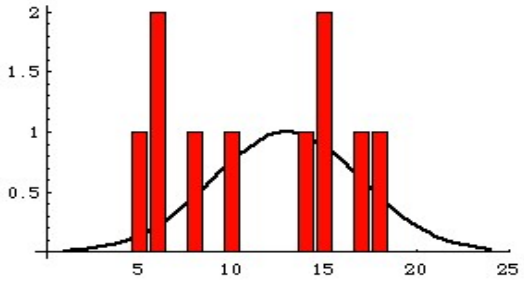


Question: Which of the following statements best describes the difference between accuracy and precision?

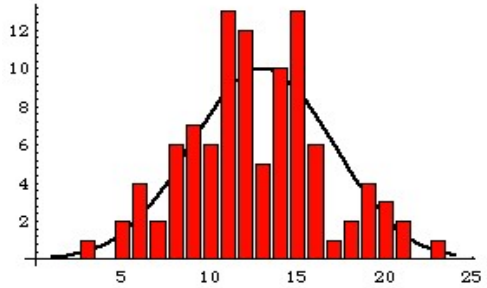
A. Accuracy refers to how close measurements are to each other, while precision refers to how close measurements are to the true value.
B. Accuracy refers to how close measurements are to the true value, while precision refers to how consistent measurements are with each other.
C. Accuracy and precision are the same and both refer to how close measurements are to the true value.
D. Accuracy and precision are unrelated to measurements and focus only on data variability.



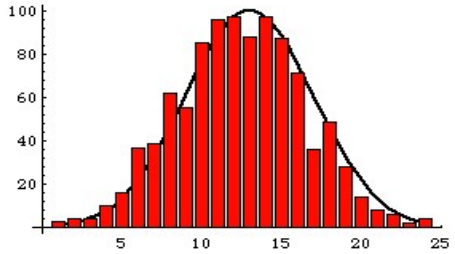




N=10



N=100



N=1000

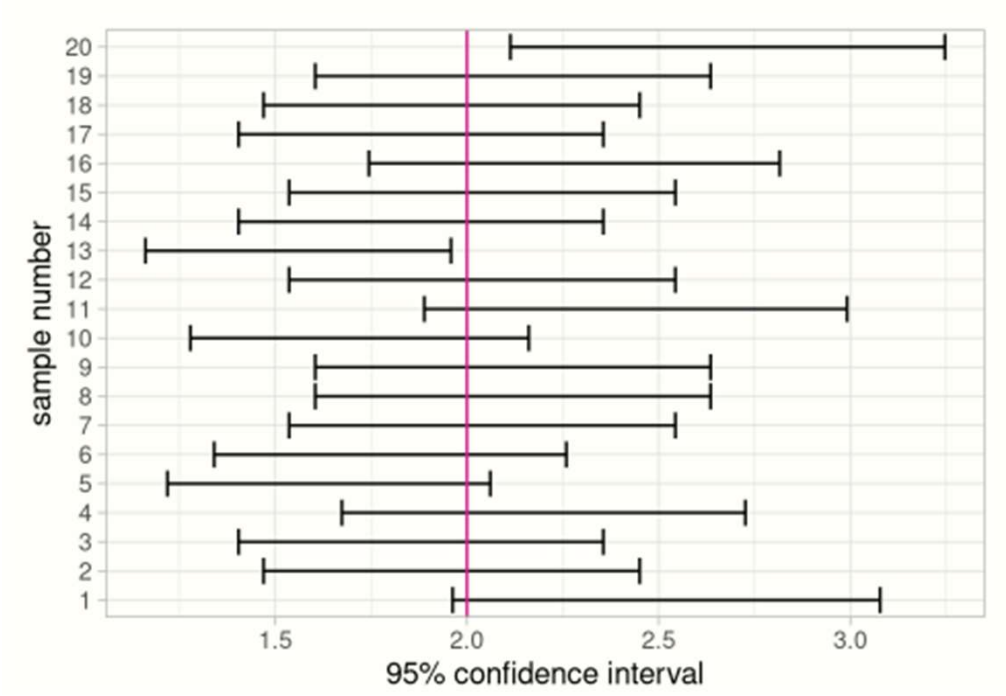
n(individual sample sizes) = 10

N is the number of repeats of sample. THIS value ranges from 10 samples to 1000 samples (each one of size 10).

95% Confidence Intervals

95% Confidence Interval is calculated:
$\bar{x} - 1.96 * SE_{\bar{x}} < \mu < \bar{x} + 1.96 * SE_{\bar{x}}$

We care a lot about precision and sample sizes because (along with alpha and some other assumptions) that is
going to create our confidence intervals!



https://www.zoology.ubc.ca/~whitlock/Kingfisher/CIMean.htm

https://stats103.com/confidence-intervals/

https://onlinestatbook.com/2/estimation/ci_sim.html

A. Can the Standard deviation ever be zero? If so, when would that situation occur?

B. Imagine that we have a population that is skewed to the left. This population has a mean of 112 and a standard deviation of 16. Using a simulation program, Tyler simulated drawing 1000 samples of size 2 from the population. He then plotted the means for each of the samples that he drew. Alex simulated drawing 1000 samples of size 30, and he also plotted the means for each of the samples that he drew.

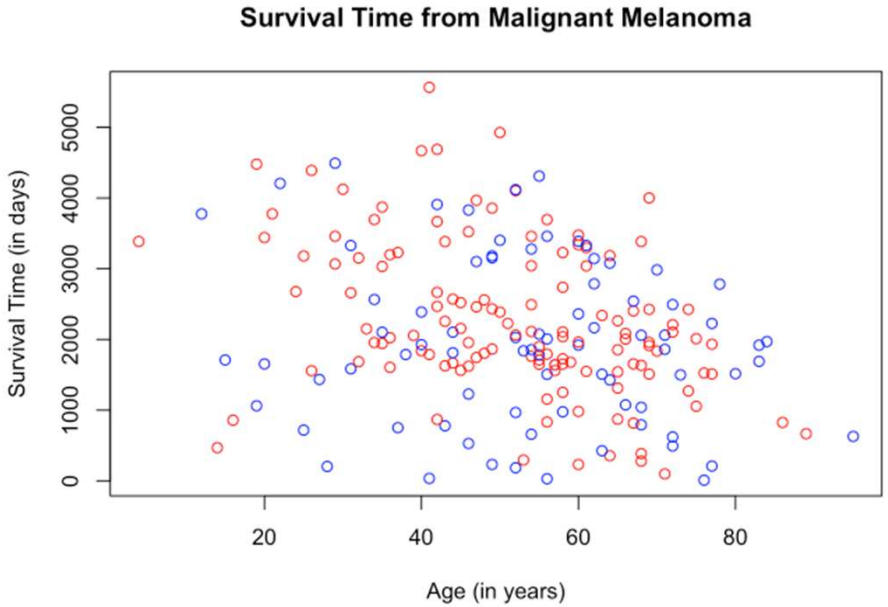
Good simulator here: **https://www.zoology.ubc.ca/~whitlock/Kingfisher/CIMean.htm**

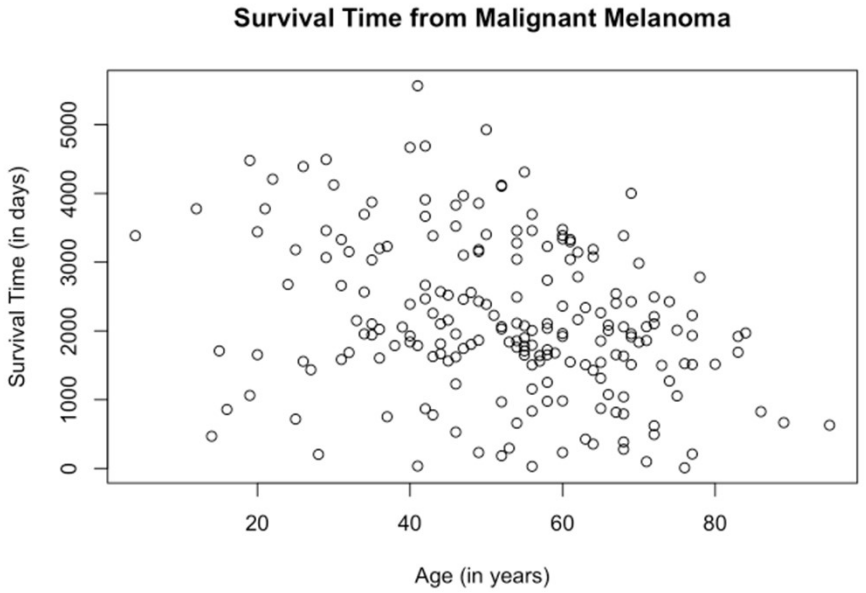
i. Would you expect the shape of Tyler' s distribution of sample means to differ from the shape of Alex's distribution of sample means? Please explain your answer (i.e., If you do expect the shapes to differ, how will they differ? If you do not expect the shapes to differ, why not?)

ii. Is the mean of Tyler's distribution of sample means <,>, or = to the mean of Alex's distribution of sample means and to the mean of the sample?

iii. How would you rank, from largest to smallest, the following: the standard deviation of Tyler's distribution of sample means, the standard deviation of Alex's distribution of sample means, and the standard deviation of the sample itself?

Scatterplot



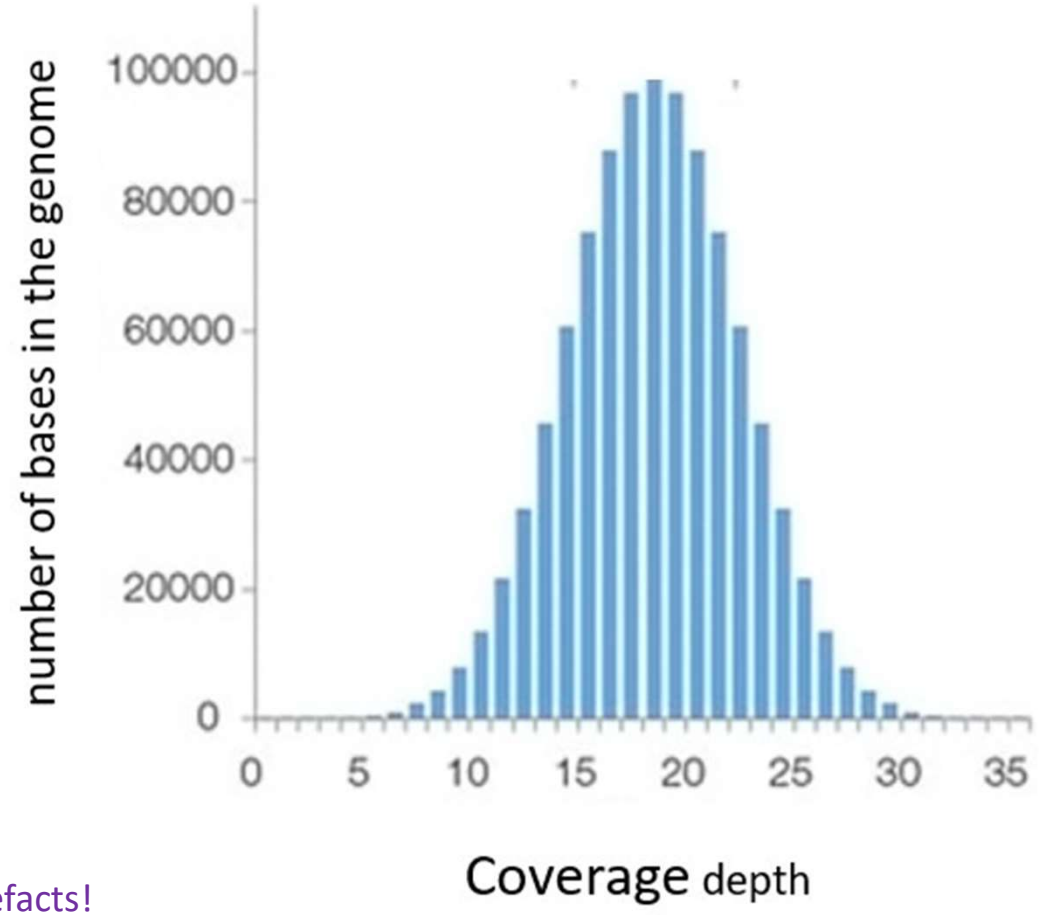


Free online textbook that gives r code!
https://bookdown.org/dli/rguide/scatterplots-and-best-fit-lines-two-sets.html

Hans Rosling ted talk (his website has data visualizations – scatterplots that move!- and datasets):
https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen

Histogram

Coverage plot of complete genome

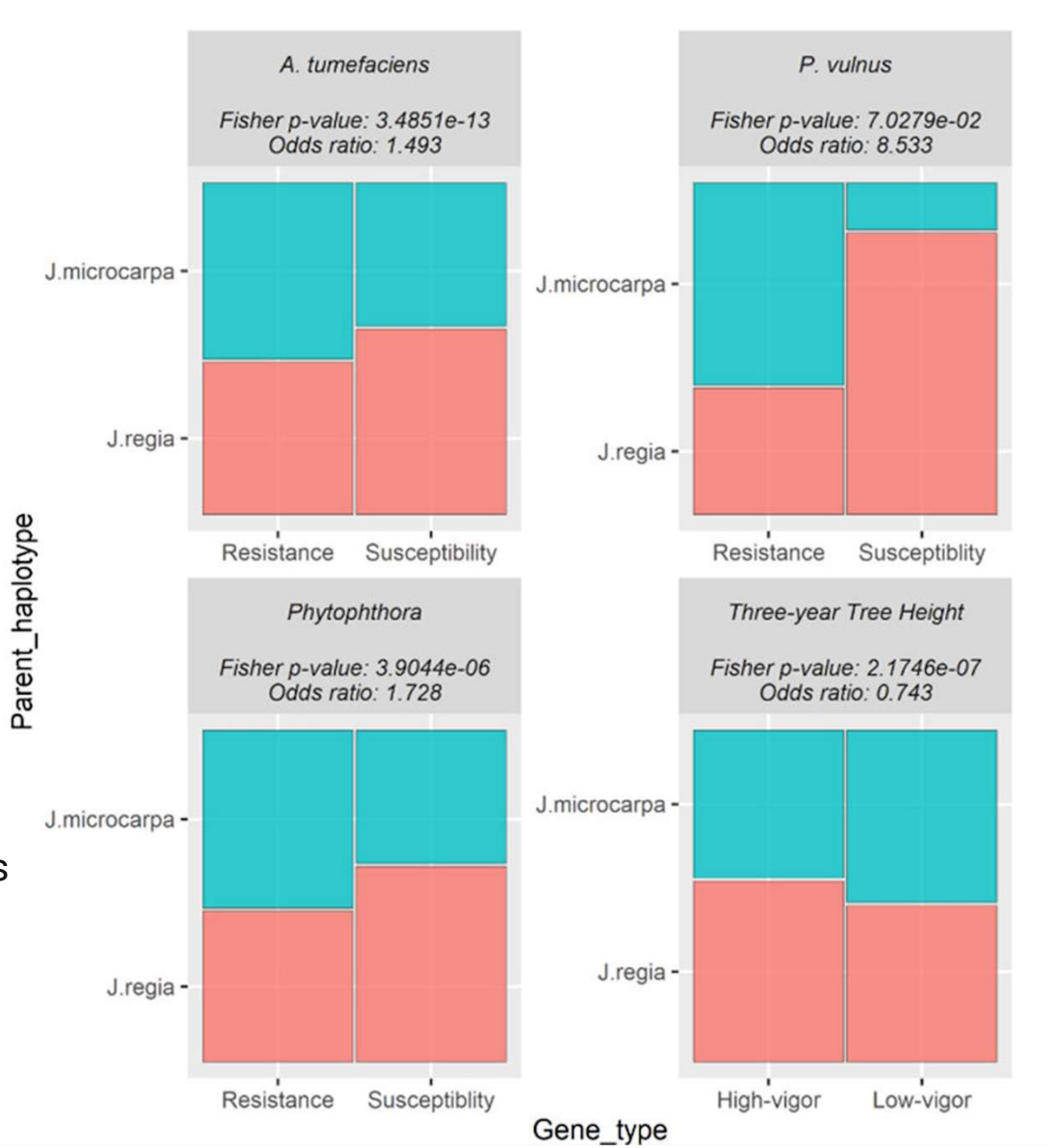


One warning about histograms:
Be careful about “bin” size; you can introduce artefacts!

https://www.biostars.org/p/9487269/

Mosaic Plot

https://www.mdpi.com/1422-0067/25/2/931

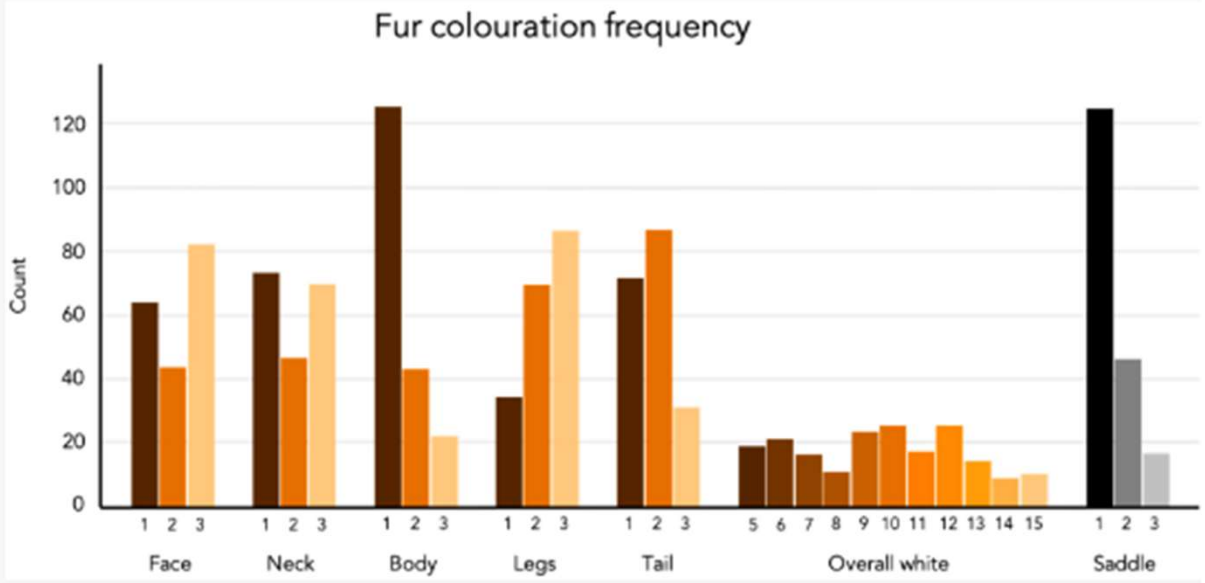


Mosaic plot representing proportions of differentially expressed genes (DEGs) from each trait colored by the haplotype the genes mapped to. Each plot is labeled with the pathogen, Fisher's exact p-value, and Fisher's exact odds ratio. The odds ratio represents the ratio of the odds of the J. regia haplotype expressing a gene positively correlated to the trait compared to the odds of the J. microcarpa haplotype expressing a gene negatively correlated to the trait.

Bar Plot

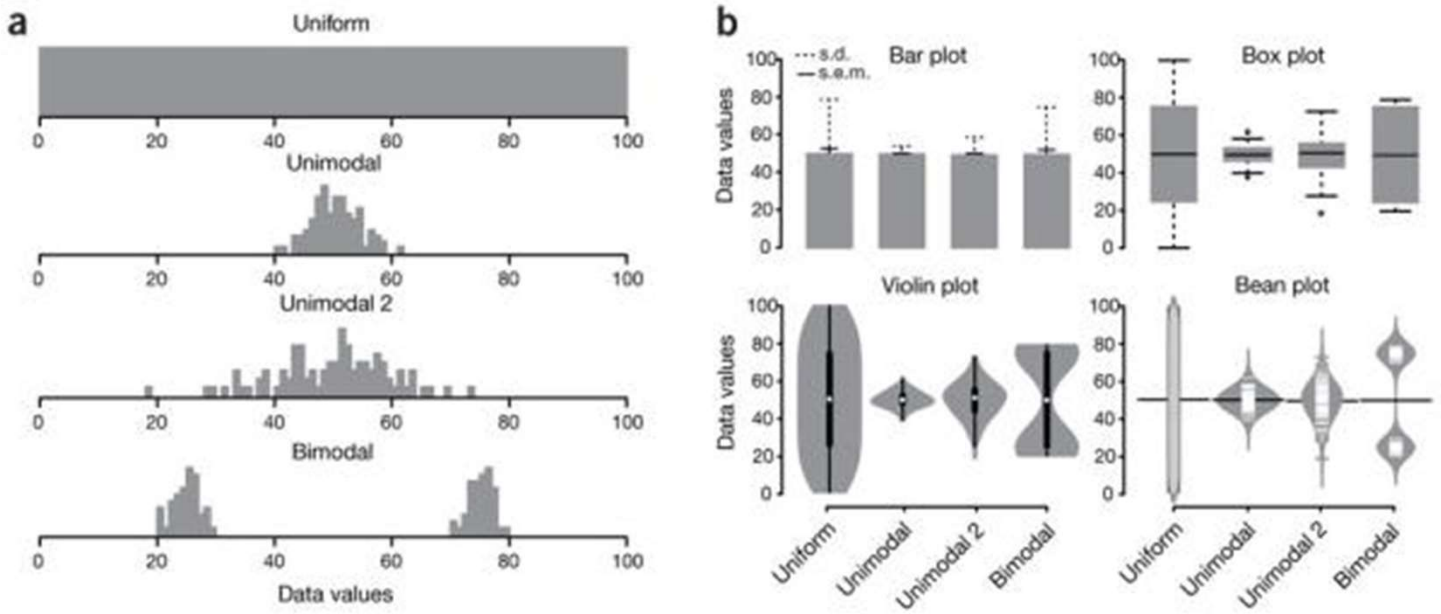
https://www.mdpi.com/2073-4425/12/2/316

Figure 1. The frequency (count) of individuals for each phenotype scoring. The total number of individuals is 190 with Table 187 due to the exclusion of three dogs that did not express the saddle.



The frequency (count) of individuals for each phenotype scoring. The total number of individuals is 190 with Table 187 due to the exclusion of three dogs that did not express the saddle

Boxplots & Violin plots



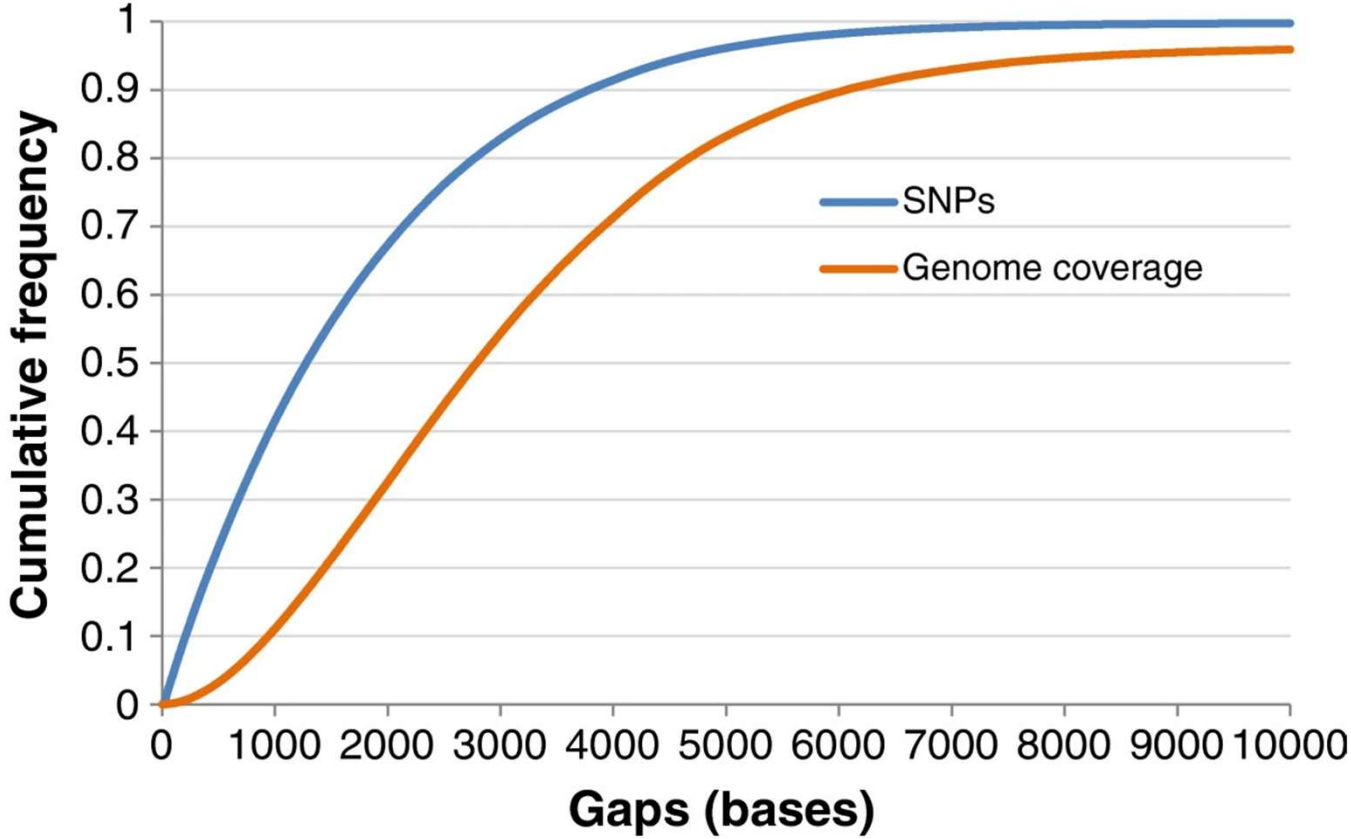
Data visualization with box plots(a)

Hypothetical sample data sets of 100 data points each that are uniform, unimodal with one of two different variances or bimodal. Simple bar plot representations and statistical parameters may obscure such different data distributions.

(b) Comparison of data visualization methods. Bar plots typically represent only the mean and s.d. or s.e.m. Box plots visualize the five-number summary of a data set (minimum, lower quartile, median, upper quartile and maximum). Violin and bean plots represent the actual distribution of the individual data sets.

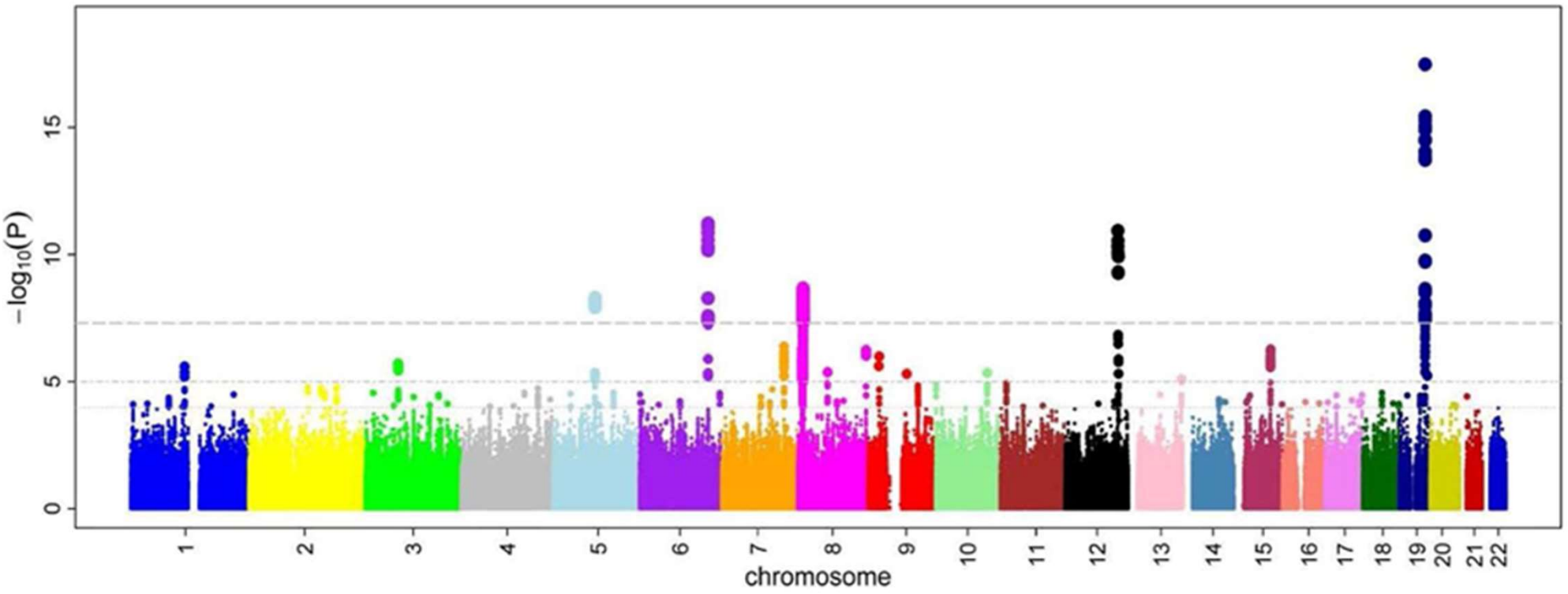
https://pubmed.ncbi.nlm.nih.gov/24481215/

Cumulative Frequency Distribution



https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-59

Cumulative frequency distributions of SNPs and genome coverage as functions of inter-marker spacing in the panel. Inter-marker spacing included distances between consecutive SNPs and the distances from chromosome ends to the nearest SNP in 600 K panel.



https://en.wikipedia.org/wiki/Manhattan_plot