

# Introduction to Statistical Inference (QTM 100 Lab)

## Lecture 4: Normal and Binomial Distributions

---

Justin Eloriaga — Emory University

Fall 2024

Data Preliminaries

Normal Distribution

Binomial Distribution

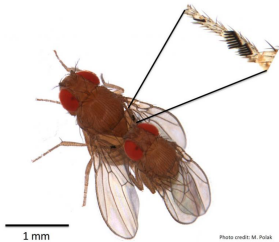
# Data Preliminaries

---

# Distributions of Random Variables

- Normal Distribution is an example of a *Continuous Random Variable*
- Binomial Distribution is an example of a *Discrete Random Variable*

# Sexual Activity and the Lifespan of Male Fruitflies, Nature, 294: 580-581 (1981)



*Drosophila bipectinata* mating pair showing the male sex comb. Matings in this species last about 10 minutes.

- Observations on 5 groups of 25 male fruitflies from an experiment designed to test if *increased reproduction reduces longevity for male fruitflies*.
- The five groups are
  1. Males forced to live alone
  2. Males assigned to live with 1 or 8 newly pregnant females
  3. Males assigned to live with 1 or 8 virgin females

# Variables Under Consideration

**No:** serial number (1-25) within each group of 25

**type:** Type of experimental assignment

1 = no females

2 = 1 newly pregnant female

3 = 8 newly pregnant females

4 = 1 virgin female

5 = 8 virgin females

**lifespan:** lifespan (days)

**thorax:** length of thorax (mm)

**sleep:** percentage of each day spent sleeping

# Importing the Dataset (again)

- Like before, we can use point-and-click or the working directory

```
setwd("YourFilePath")
```

```
fruitfly <- read.csv("fruitfly.csv", header = TRUE)
```

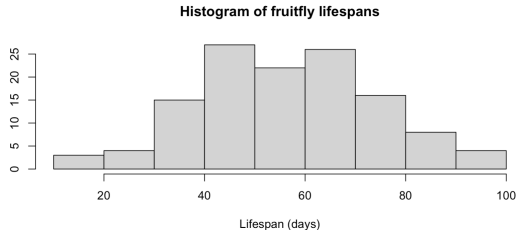
- Let's also examine the structure and give an overview of the dataset

```
summary(fruitfly$lifespan)
```

# Visualizing Fruitfly Lifespans

We can generate a histogram for fruitfly lifespans using

```
hist(fruitfly$lifespan, main = "Histogram of Fruitfly Lifespans", xlab = "Lifespan (days)", ylab = "")
```



The lifespan of fruitflies ranges from 16 to around 97 days with an average of 57.4 days.

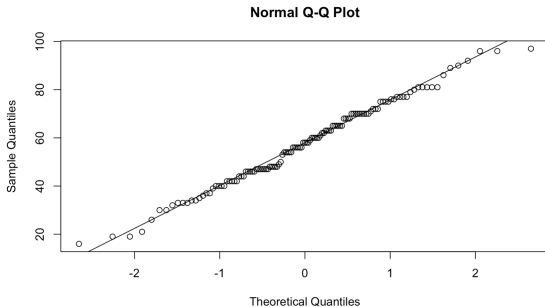
Is this distribution normal?



# QQ Plots

To understand better whether or not the distribution is normal, we can use a QQ Plot

```
qqnorm(fruitfly$lifespan)  
qqline(fruitfly$lifespan)
```



What do you notice?

# Normal Distribution

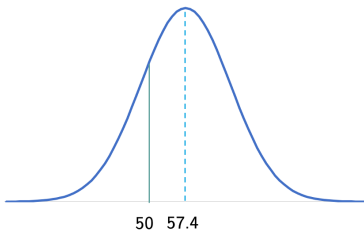
---

# Answering Probability Questions

- Once you have identified that a variable is approximately normal, we can now calculate probabilities regarding outcomes of this variable.
- For example, what is the probability that a fruitfly lives less or equal to 50 days?
- We need two other ingredients!
  1. mean
  2. standard deviation

# Using `pnorm`

The `pnorm` command gives us the area under the normal curve given some mean and std. deviation



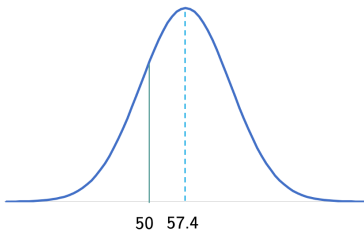
- Run `mean()` and `sd()` commands to get the mean and standard deviation.
- We use the command below to calculate the area under the curve

```
pnorm(q = 50, mean = 57.4, sd = 17.6)
```

- Based on the `pnorm` command, 33.7% of fruitflies are expected to survive less than or equal to 50 days!

## Using `pnorm`

The `pnorm` command gives us the area under the normal curve given some mean and std. deviation



- Run `mean()` and `sd()` commands to get the mean and standard deviation.
- We use the command below to calculate the area under the curve

```
pnorm(q = 50, mean = 57.4, sd = 17.6)
```

- Based on the `pnorm` command, 33.7% of fruitflies are expected to survive less than or equal to 50 days!

**QUESTION:** How does this compare to the actual data?

## Close enough.. HAHA

To compare our result using `pnorm` (i.e. the z-score and the z-table), we compute using the command below

```
sum(fruitfly$lifespan <= 50)/length(fruitfly$lifespan)
```

- `sum()` counts how many observations survived less than or equal to 50 days
- `length()` gives us the sample size
- Based on our data, our empirical estimate is that 39.2% survived less or equal to 50 days! Very close!

**Aside:** If you want the upper tail, just do 1 minus what we computed for

```
1 - pnorm(q = 50, mean = 57.44, sd = 17.56389)
```

# On Intervals

Example: What is the probability that a fruitfly survives between 50 and 70 days?

```
pnorm(q = 70, mean = 57.4, sd = 17.6) - pnorm(q = 50, mean = 57.4, sd = 17.6)
```

This gives us that 42.5% of fruitflies survive between 50 and 70 days.

**Aside:** The above operation can also be operationalized using the `diff()` function

```
diff(pnorm(q = c(50,70), mean = 57.4, sd = 17.6))
```

## More on `pnorm()` and `qnorm()`

We can use the normal distribution to find percentiles of the normal using `qnorm()` where `p` is the percentile of interest.

Example: What is the value of the 90th percentile? A: 79.95531

```
qnorm(p = 0.9, mean = 57.4, sd = 17.6)
```

Example: 79.95531 is what percentile? A: 90th percentile!

```
pnorm(q = 79.95531, mean = 57.4, sd = 17.6)
```

The two commands are very related! Based on the normal distribution, 90% of fruitflies survive approx. 80 days or less.



# Binomial Distribution

---

# Motivating the Binomial Distribution

We know that the prob. a fruitfly survives more than 50 days is 0.663. Suppose we are experimenting on a new group of 8 fruitflies and want to know the likelihood that a certain number of them survive more than 50 days.

We can use the BINOMIAL DISTRIBUTION!

1. Fruitflies are assumed independent
2. Fixed number of fruitflies
3. Each fruitfly can either survive more than 50 days or die before or at 50 days
4. Each fruitfly is assumed to be equally likely to survive

## Using `dbinom()`

Using `dbinom()` gives the probability that a certain number of fruitflies survive more than 50 days.

For example, the probability that exactly 5 out of 8 fireflies survive more than 50 days is calculated by

```
dbinom(x = 5, size = 8, prob = 0.663)
```

Ergo, 27%.

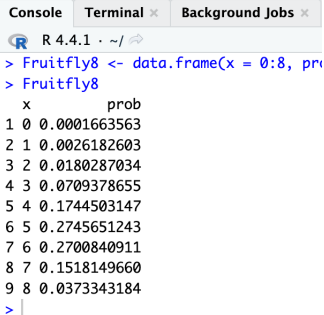
- `x` is the number of survive fruitflies (number of successes),
- `size` is the total number of fruitflies (number of trials)
- `prob` is the prob. to survive more than 50 days (prob of success)

To compute for multiple outcomes, use `dbinom(x = 0:8, size = 8, prob = 0.663)`

# Using a data frame

Using a data frame allows us to see the results more clearly

```
Fruitfly8 <- data.frame(x = 0:8, prob = dbinom(x =  
0:8, size = 8, prob = 0.663))  
Fruitfly8 # Look at the Data Frame Created
```



The screenshot shows an R console window with three tabs: Console, Terminal, and Background Jobs. The Console tab is active, displaying the R prompt and the execution of the following commands:

```
> Fruitfly8 <- data.frame(x = 0:8, prob = dbinom(x =  
> Fruitfly8
```

The output is a data frame with two columns: 'x' and 'prob'. The values for 'x' are 0 through 8, and the values for 'prob' are the corresponding binomial probabilities calculated by the dbinom function.

	x	prob
1	0	0.0001663563
2	1	0.0026182603
3	2	0.0180287034
4	3	0.0709378655
5	4	0.1744503147
6	5	0.2745651243
7	6	0.2700840911
8	7	0.1518149660
9	8	0.0373343184

- First column will have a name x with numbers from 0 to 8
- Second column will have a name prob with the values we computed for

## Range of Values in a Binomial Setting

Using the `sum()` command is very useful!

Note: prob. that less than or equal to 5 fruitflies survive more than 50 days = prob. that between 0 and 5 fruitflies survive more than 50 days

Notice that these two commands yield the same answer

```
sum(dbinom(x = 0:5,size = 8, prob = 0.663))
```

```
sum(Fruitfly8$prob[Fruitfly8$x <= 5])
```