WEEK 2, LOTS OF EXERCISES!

SCHOOL OF
CULTURE AND SOCIETY
AARHUS UNIVERSITY

23 JANUARY 2022

SIGURD FYHN SØRENSEN
STUDENT TEACHER

PLAN FOR TODAY.

- Recap
- Expected Value
- Linear transformations
- Probability distribution
- Finding Standard Error for Binomial Data.

- Lots of time for exercises to get your coding fingers going again.

RECAP.

1. Why do we call it the General Linear model?
2. How do I create my own version of a github repository?
3. How do I get a local version of a github repository.
4. How do I pull?
5. How do I push?

All the steps please. ☺

EXPECTED VALUE/AVERAGED MEANS:

- Arithmetic mean:
  - $A = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + a_3 \ldots + a_n}{n}$
- Males' height = [180,183,190,179,184,182,180], Females' height = [160,170,173,191]

```{r}
male <-  c(180,183,190,179,184,182,180)
female <-  c(160,170,173,191)

height <- c(male,female)
mean_1 <- mean(height)
sem_1 <- sd(height)/sqrt(length(height))


sem_1 ; mean_1
```

```
 [1] 2.686975
 [1] 179.2727
```

- But… Sometimes we don't have the underlying distributions.

EXPECTED VALUE/AVERAGED MEANS:

- If we don't have the underlying data points. But...
  - N_female = 4, N_male = 7
  - $\bar{y}_{female} = 173.5$, $\bar{y}_{male} = 182.5714$
- We can use $$\text{weighted average} = \frac{\sum_j N_j \bar{y}_j}{\sum_j N_j},$$
- $\frac{4*173.5 + 7 * 182.5714}{4+7} = 179.2727$

```{r}
(length(male)*mean(male)+length(female)*mean(female)) / (length(female)+length(male))
```

```
[1] 179.2727
```

EXPECTED VALUE/AVERAGED MEANS:

- From Weighted Average to Expected Value.
  - Expected value is denoted as E[X] and its generalization of the averaged means.
  - Now using proportions/probabilities instead of counts.

$$\mathrm{E}[X] = \sum_{i=1}^{\infty} x_i\, p_i \qquad \longleftrightarrow \qquad \text{weighted average} = \frac{\sum_j N_j \bar{y}_j}{\sum_j N_j},$$

$$\mathrm{E}[X] = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$

```{r}
mean_female <- mean(female)
mean_male <- mean(male)

n <- length(male) + length(female)
p_male <- length(male)/n
p_female <- length(female)/n

1 == p_male + p_female

p_male*mean_male + p_female*mean_female
```

- $p_i = \frac{n_i}{\sum p_i}$, $\qquad P = \sum \frac{n_i}{\sum_{i=1}^{n} p_i} = 1$

- **<u>This can be useful</u>** when you are not given the underlying datapoints. Much data is given like this.

STANDARD DEVIATION OF BINOMIAL DATA

- A person drinks 30 beers and their probability of vomiting is p = 0.3 with each beer.
  - We assume that the probability of vomiting is independent (questionable).
  - Usually, repeated observations are depended (mixed effect models) but let's keep it simple.
- **Expected value**. E[pukes] = 0.3 * 30 = 9. There is of course some uncertainty.

- SD_binom = $\sqrt{np(1-p)}$ = $\sqrt{30 * 0.3\ (1-0.3)}$ = 2.51

SE = $\sqrt{p(1-p)/n}$ = $\sqrt{0.3 * \frac{0.7}{30}}$ = 0.086

95% Confidence interval = $[0.3 \pm 1.96$ * $\sqrt{0.3 * \frac{0.7}{30}}\,]$

ANOTHER INSTANCE

- We want to find the estimate the difference between men and woman puking when drinking beer based on our sample.
- **<u>Sample:</u>** 1000 people drink bears. 400 men and 600 women. 62% of the men puke and 47% of the women puke.

- **<u>The standard errors for these proportions</u>**
- SE = $\sqrt{p(1-p)/n}$
- SE_women = $\sqrt{\frac{0.47 * 0.53}{600}} = 0.02$, SE_men = $\sqrt{\frac{0.62*0.38}{400}} = 0.024$

<u>Estimated gender gap between people who puke:</u>

- 0.62 – 0.47 = 0.15 with standard error…
- SE_diff = $\sqrt{se_{men}^2 + se_{women}^2} = \sqrt{0.02^2 + 0.024^2} = 0.031$

LINEAR TRANSFORMATION

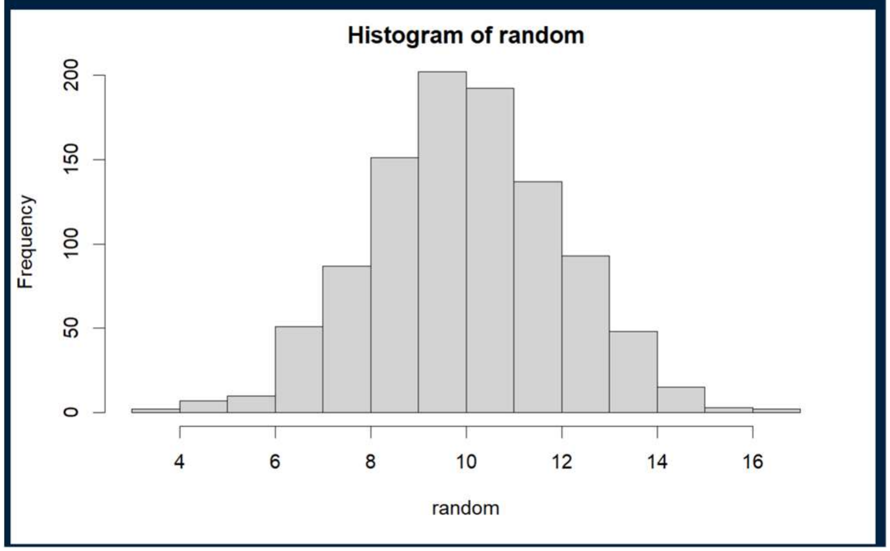
- A linear transformation changes the original value of x into a new variable x-new.
  - Without changing the overall relationship.

$$x_{new} = a + bx$$

```{r}
sd(2*random)
mean(2*random)
```

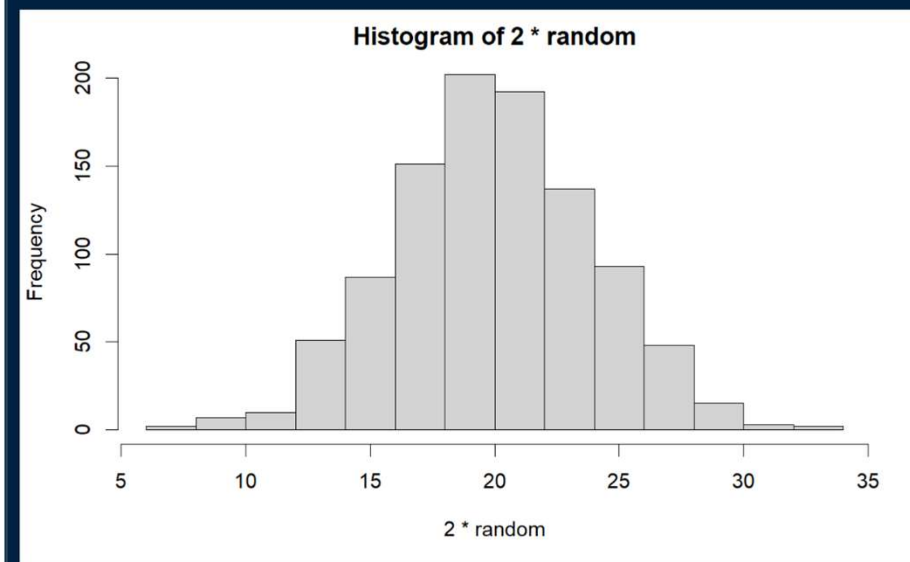
```
[1] 4.00461
[1] 19.9735
```

```{r}
random <- rnorm(1000, mean = 10 , sd = 2)
hist(random)
```



2*

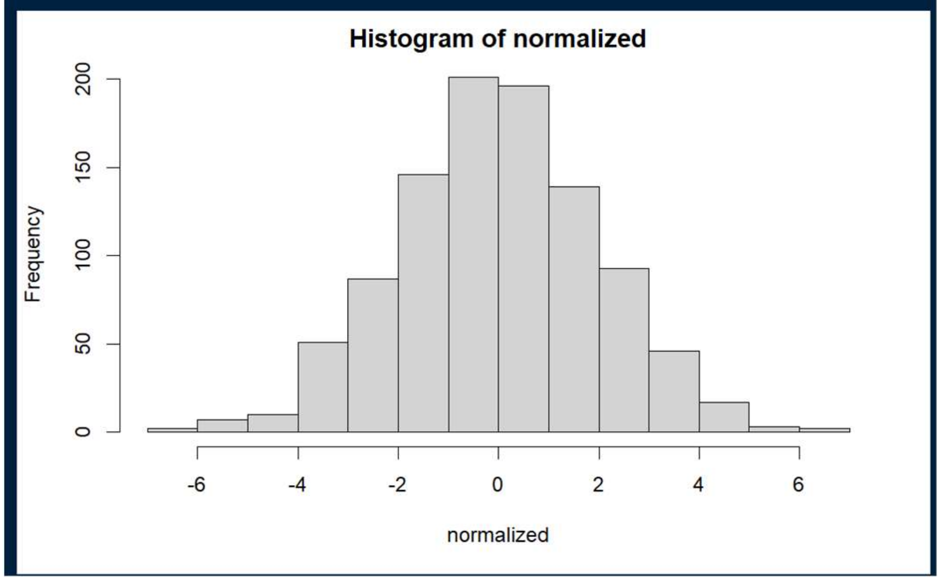
```{r}
hist(2*random)
```



NORMALIZED

- So, we can multiply but we can also add.
  - A special form of linear transformation is when you normalize and standardize.

```
```{r}
normalized <- random-mean(random)
hist(normalized)
```
```

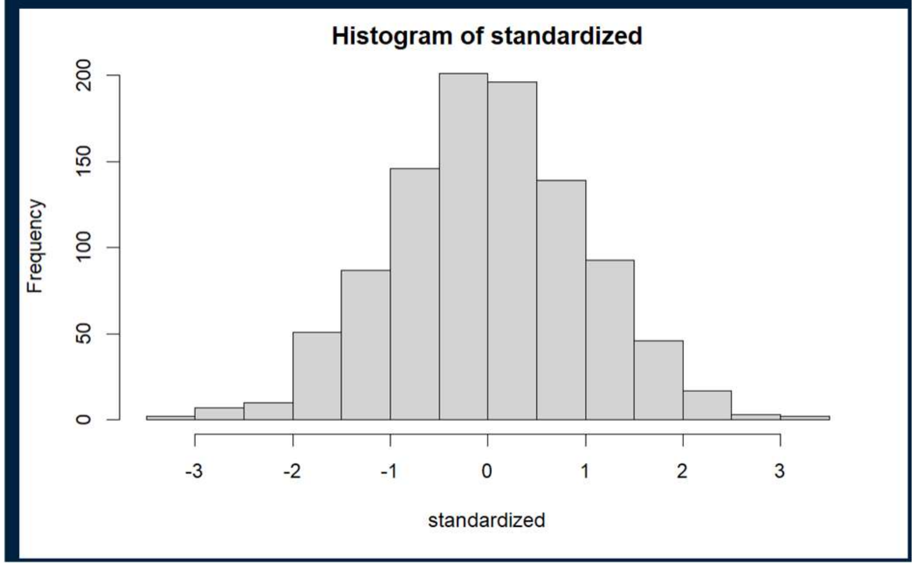


Measure	x	x_{new}
Mean	$\bar{x}$	$a + b\bar{x}$
Median	M	a+bM
Mode	Mode	a+bMode
Range	R	$\lvert b \rvert R$
IQR	IQR	$\lvert b \rvert IQR$
Stdev	s	$\lvert b \rvert s$

STANDARDIZED

- When we standardize, we want a mean = 0 and sd = 1
- Using: $y_i = m2 + (x_i - m1) \times s2/s1$

```{r}
standardized <- normalized*1/sd(normalized)
hist(standardized)
```

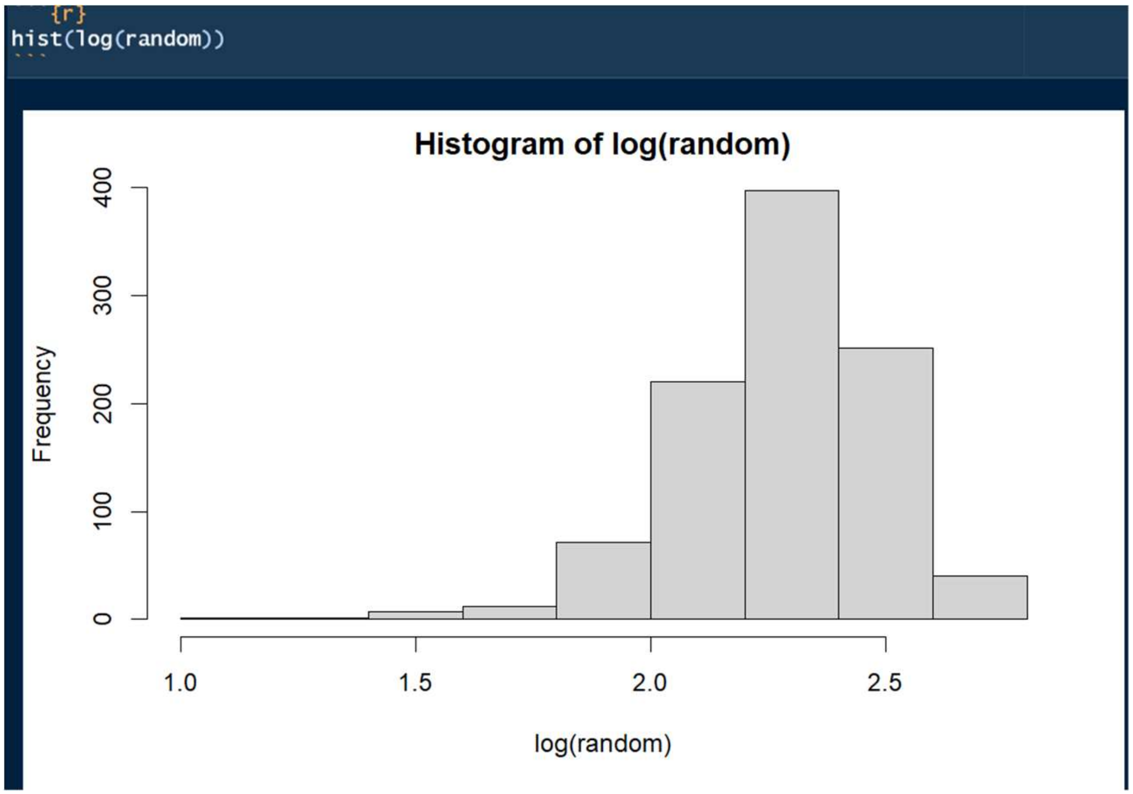



```{r}
sd(standardized)
mean(standardized)
```

```
[1] 1
[1] 1.653498e-16
```

In all the cases the relation between the observations remain the same. The distribution is the same and the proportional distance between observations are the same.

NON-LINEAR TRANSFORMATION



- The resulting X-new no longer has the relative relation between observations.

- Log() is not a type of linear transformation
    - Non-linear transformation is also useful, but we will get to that.

- JUST wait for the math weeks!

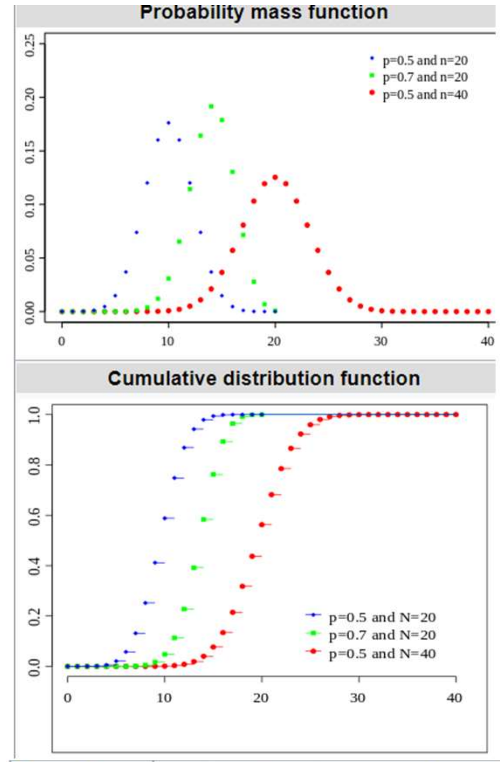
PROBABILITY DISTRIBUTIONS

- Probability distribution allows us describe X and its variability.
- Frequency plots (histogram) good but probability distributions integrate to 1.
  - Useful to describe the spread of the data.
  - Likelihood of the different observations.

- Distribution can be gaussian data.
  - But not required.
  - Binomail and Poisson is also possible.

Let's look in R and at some binomial distribution.

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



Parameters	$n \in \{0, 1, 2, \ldots\}$ – number of trials $p \in [0, 1]$ – success probability for each trial $q = 1 - p$
Support	$k \in \{0, 1, \ldots, n\}$ – number of successes
PMF	$\binom{n}{k} p^k q^{n-k}$

QUICK FUNCTION.

Rpois() to sample from a poisson distribution.

Rbinom() to sample from a binomial distribution.

Rnorm() to sample from a gaussian distribution.

They all come with their supplementary functions, see ?rbinom() ?rpois() & ?rnorm() or google https://www.statology.org/dbinom-pbinom-qbinom-rbinom-in-r/

EXERCISES:

Easy/Medium (must do)
3.1 3.3 3.4 3.5 & 3.6
4.1 4.2 4.3 4.4 & 4.5

Use the skills you learned:
Exercise 3.10 & 4.11

Hard: (challenge)
3.2 3.8 & 4.7

AARHUS
UNIVERSITY