Problem Set 3 Key

1 Problem 1

The table below shows the frequency table of the number of referrals offered for expectant mothers who experience socioeconomic barriers for needed healthcare.

Number of Referrals	Frequency
0	90
1	132
2	76
3	10

- What is the probability that a randomly selected participant received at least one referral? [2 pts.]

> Note
>
> We need to calculate the relative frequencies for each category first to calculate the specific probabilities.

```r
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.5.1
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.0     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
```

```
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
df <- data.frame(Referrals=c(0,1,2,3),
                 Frequency = c(90,132,76,10))

mutate(df,RelFrequency = Frequency/sum(Frequency))
```

```
  Referrals Frequency RelFrequency
1         0        90   0.29220779
2         1       132   0.42857143
3         2        76   0.24675325
4         3        10   0.03246753
```

OR

```r
total <- sum(c(90,132,76,10))

df <- data.frame(Referrals=c(0,1,2,3),
                 Frequency = c(90,132,76,10),
                 Rel.Frequency = c(90,132,76,10)/total)

df
```

```
  Referrals Frequency Rel.Frequency
1         0        90    0.29220779
2         1       132    0.42857143
3         2        76    0.24675325
4         3        10    0.03246753
```

$$P(X \geq 1) = P(X = 1) + P(X = 2) + P(X = 3)$$

or

$$P(X \geq 1) = 1 - P(X = 0)$$

Hence the probability $P(X \leq 1)$ is:

```r
1-90/sum(c(90,132,76,10))
```

```
[1] 0.7077922
```

```r
# OR

1-0.2922
```

```
[1] 0.7078
```

- What is the expected value of the number of referrals in the sample? [1pt.]

> ℹ Note
>
> The expected number of referrals is the sum of the product of the outcome and their corresponding probabilities.
>
> $$E(X) = 1 * (90/308) + 2 * (132/308) + 3 * (76/308) + 4 * (10/308)$$

```r
0*(90/308) + 1*(132/308) + 2*(76/308) + 3*(10/308)
```

```
[1] 1.019481
```

```r
# OR

sum(df$Referrals*df$Rel.Frequency)
```

```
[1] 1.019481
```

2 Problem 2

The proportion of individuals from a certain population with an O blood type is 0.44. What is the probability that out of a sample of 30 individuals from this population,

> ℹ Note
>
> We use the binomial distribution!

a. That exactly half of the individuals in the sample have an O blood type?

```r
dbinom(15,30,0.44)
```

```
[1] 0.1162175
```

b. At most 10 individuals have an O blood type? [1pt]

```r
pbinom(10,30,0.44)
```

```
[1] 0.160397
```

3 Problem 3

```
If the mean number of unexcused absences per school year for fifth graders at a public elemen
```

> **Note**
>
> We use the Poisson distribution!

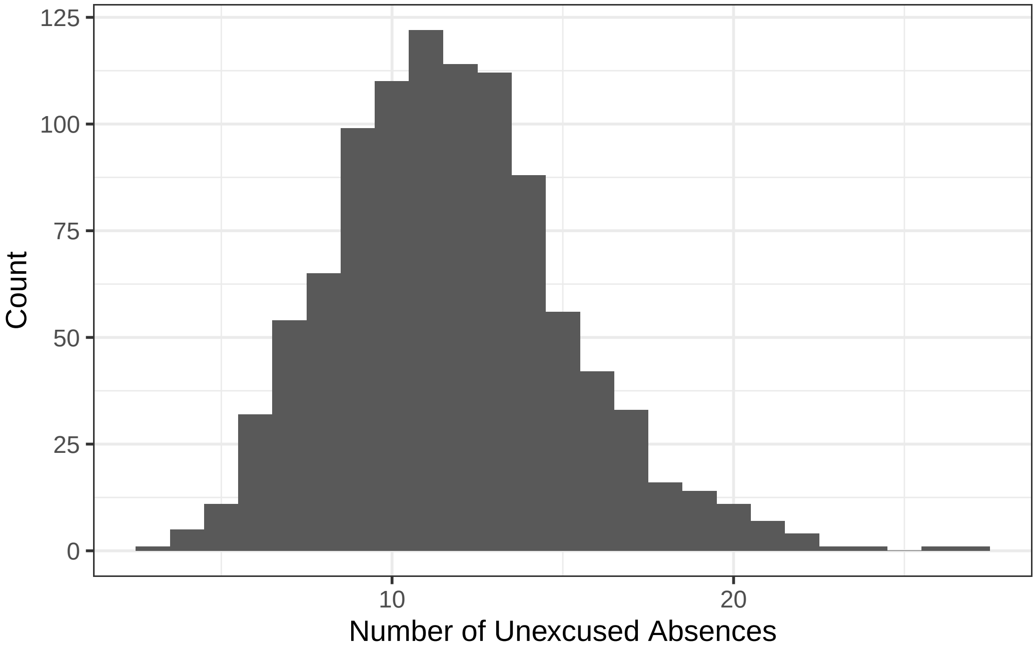
a. What is the probability that a randomly selected student will have chronic absenteeism, i.e. at least 18 unexcused absences per school year? [1 pt.]

```r
1-ppois(17,lambda=11.8)
```

```
[1] 0.05561775
```

b. Generate 1000 numbers from the associated distribution and create a histogram from these values. [2pts.]

```r
library(tidyverse)
Nsim<- 1000 # Simulating 10 times
set.seed(12)
sims <- rpois(Nsim,11.8)
df <- data.frame(simulations=sims)
ggplot(df, aes(x=simulations)) +
  geom_histogram(binwidth=1) +
  theme_bw() +
  labs(x="Number of Unexcused Absences", y="Count")
```



4 Problem 4

A random number generator was used to assign participants to two treatment groups. A random number was generated from a UNIF(-1,1) distribution. A value between -0.3 and 0.2 means the participant is assigned to treatment group A, else they were assigned in B. What is the probability that a participant is assigned in group A? [2pts.]

> **Note**
>
> You can use R or manual mathematical calculation of areas. I will be showing the R solution.
> We use the CDF to calculate probabilities for continuous distributions such as the normal distribution. We will be using `pnorm()`.

```
punif(0.2,-1,1) - punif(-0.3,-1,1)
```

```
[1] 0.25
```

5 Problem 5

The mean A1c measurement for residents at a nursing home was 5.6 with a standard deviation of 2.1. Assuming the A1c measurements follow a normal distribution,

> Tip
>
> We use the CDF to calculate probabilities for continuous distributions such as the normal distribution. We will be using `pnorm()`.

a. What is the probability of randomly selecting a resident who is prediabetic (A1c between 5.7 and 6.4)? [1pt]

```
pnorm(6.4,mean=5.6,sd=2.1) - pnorm(5.7,mean=5.6,sd=2.1)
```

```
[1] 0.1293906
```

b. What is the probability of randomly selecting a resident who is diabetic (A1c above 6.5)? [1pt]

```
pnorm(6.5,mean=5.6,sd=2.1, lower.tail = F)
```

```
[1] 0.3341176
```

c. What is the expected number of diabetics out of a sample of 100 residents based on this distribution? (Hint: Randomly selecting a diabetic in this sample is a Bernoulli process.) [2 pts.]

> Tip
>
> Remember that for a Bernoulli process, the expected value $E(X) = np$ where n is the sample size and p is the probability of success.

```
probability_diabetic <- pnorm(6.5,mean=5.6,sd=2.1, lower.tail = F)
sample_size <- 100

sample_size*probability_diabetic
```

```
[1] 33.41176
```

Expected value is `r sample_size*probability_diabetic`. If we're talking about number of diabetics, we **round up** to 34.

d. What is the 85th percentile of the A1c measurements in this nursing home? [1 pt.]

> Tip
>
> We can use `qnorm()` to calculate values based on percentiles.

```r
qnorm(0.85,mean=5.6, sd=2.1)
```

```
[1] 7.77651
```