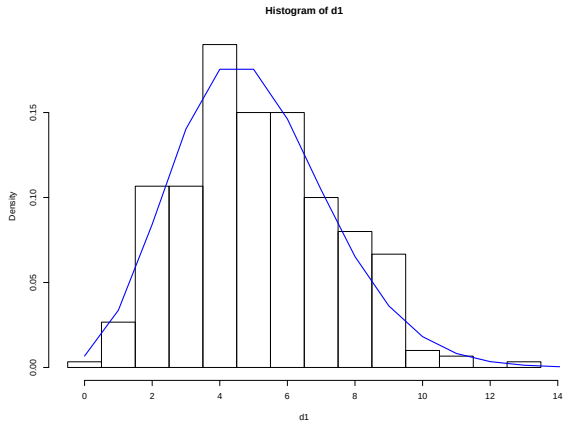
R Lab: Fitting probability distributions to data

SOLUTIONS.

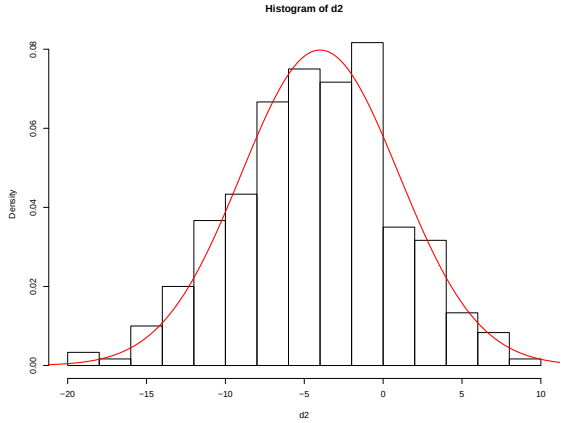
Part I. Fitting data.

The dataset $d1$ is best modeled by a ????? random variable with parameters ????:

1. $D1 \sim$ Pois(5) $D1$ might be used to model ..... the number of phone calls received in a one hour period.

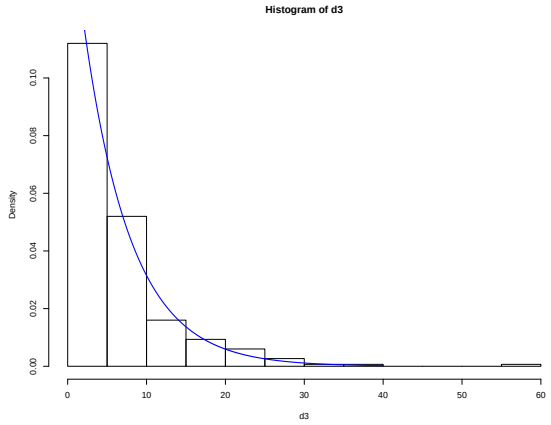


2. $D2 \sim$ Norm(-4, 5). This means $\mu = -4$ and $\sigma^2 = 5$. $D2$ might be used to model ..... the average per annum change in ten of thousands of dollars of a small business' profits. WARNING post-grading: When giving the parameters for a normal distribution, give the variance σ^2.



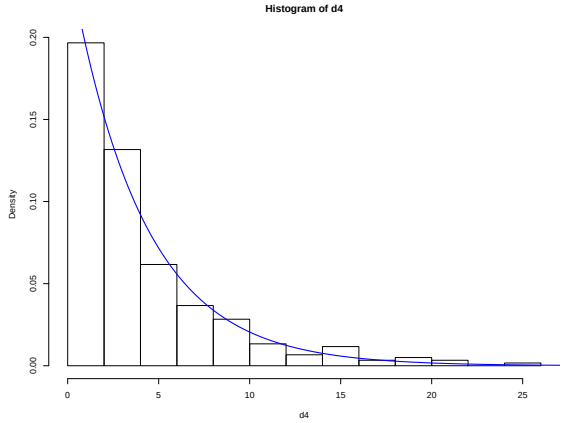
3. $D3 \sim$ Exp(6). That is, $\alpha = 1$, $\beta = 6$.

$D3$ might be used to model ..... the waiting time in years until the next failure of an LED lightbulb.



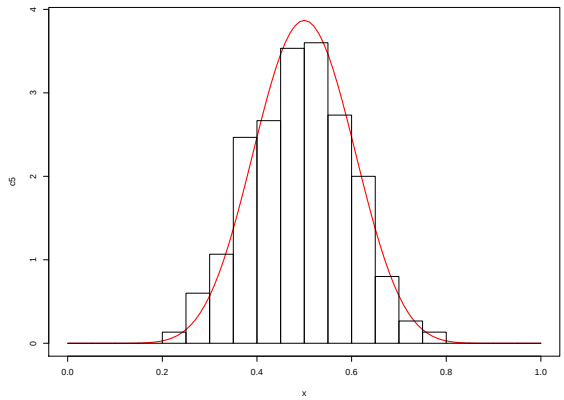
4. $D4 \sim$ Exp(4). That is, $\alpha = 1$, $\beta = 4$.

$D4$ might be used to model ..... the length of time in years until the next crack in your car's windshield.



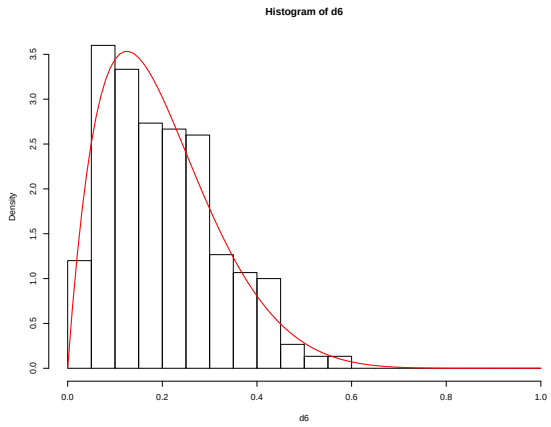
5. $D5 \sim$ Beta(12,12)

$D5$ might be used to model ..... an estimate for the *unknown* probability that a H comes up when a coin is flipped, when in truth the coin is fair. Comments after grading: The Beta distribution is a *better* fit than a normal since the values are confined to the unit interval $[0, 1]$.



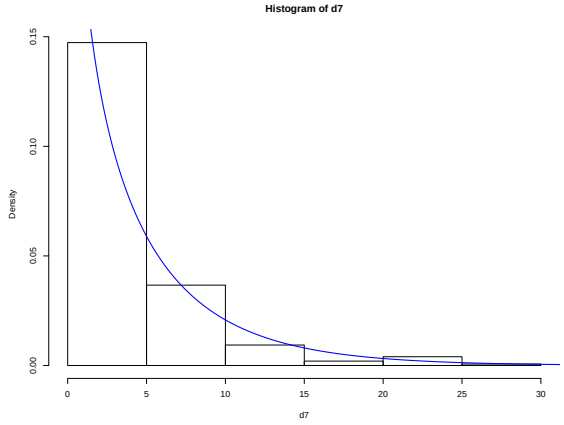
6. $D6 \sim$ Beta(2,8)

$D6$ might be used to model ..... the proportion of one hour spent by workers cleaning your car at the car wash.



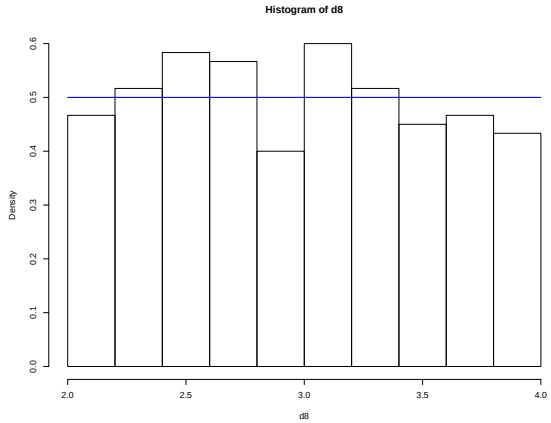
7. $D7 \sim$ Gamma(.7,6)

$D7$ might be used to model ..... the number of minutes kept on hold when calling a credit card company with a complaint.



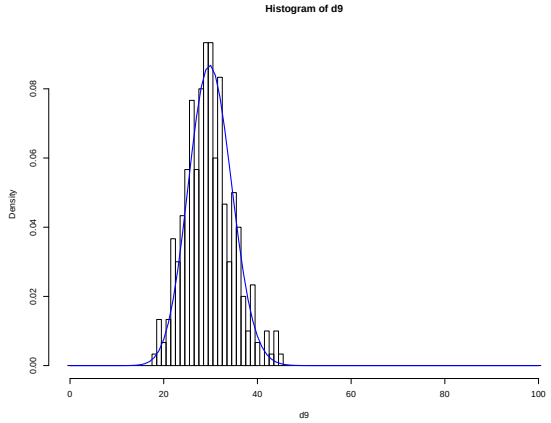
8. $D8 \sim$ Unif(2,4)

$D8$ might be used to model ..... the number of hours spent studying for a Probability exam in a population of college students.



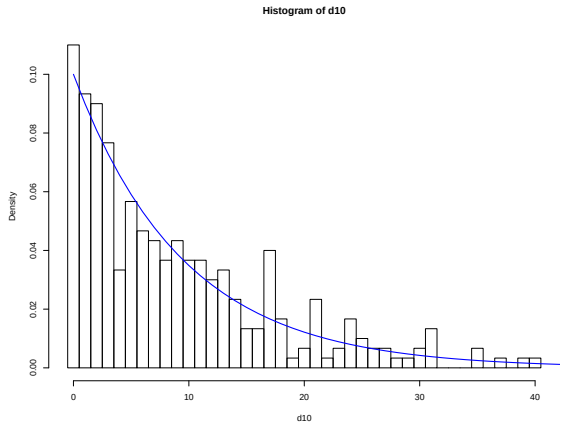
9. $D9 \sim$ Binom(100,.3)

$D9$ might be used to model ..... the number of people who favor the death penalty in a sample of 100.



10. $D10 \sim$ Geom(.1)

$D10$ might be used to model ..... the number of lightbulbs tested [minus one] until a defective is found. (Recall that in R a geometric random variable counts the trial on which the last failure occurs, hence the positive probability for the outcome 0. This was mentioned in class in response to a student question.)

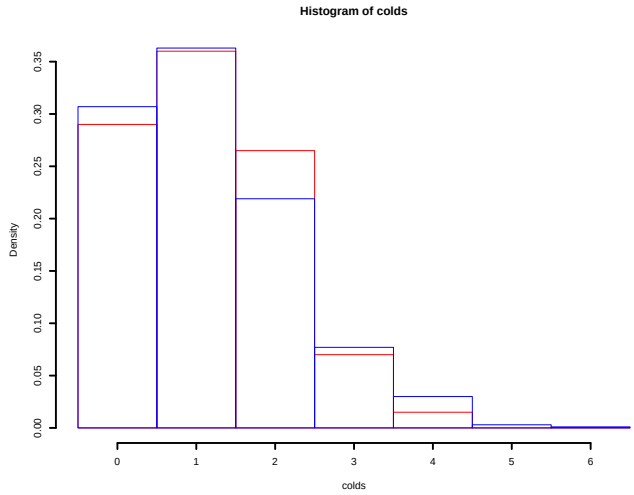


Part II. Data on students collected from a sample of size 200.

Answers:

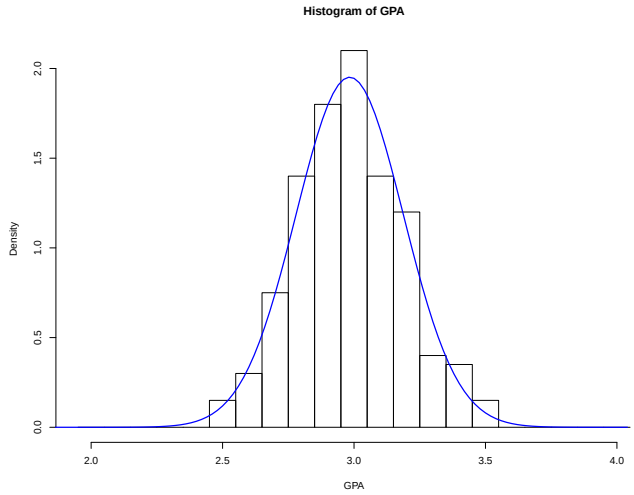
1. The variable `colds` is best modeled by a Poisson random variable with $\lambda = 1.16$ because ... colds are a rare event in a year.

   I chose the value of $\lambda = 1.16$ by using the `summary` command in R and finding that this was the mean. The histogram for the data is shown in red, and for the Poisson fit in blue. To make the Poisson histogram, I drew 1000 random samples from Pois(1.16) and used these to make the blue histogram. Comments after grading: A number of students suggested that this be modeled with a Geometric random variable. This is not a good choice, as $X \sim \text{Geom}(p)$ is a variable that counts the trial on which the first success occurs. That is not the case for the variable Colds. Choosing a Binomial variable is also a poor choice here for a similar reason.



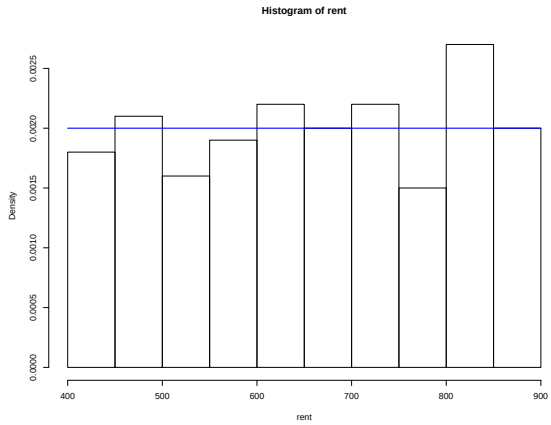
2. The variable `GPA` is best modeled by a Norm(2.9,.0417) random variable (or something close. I suspect I generated it with $\mu = 3$ and $\sigma^2 = .04$.

   because one might expect GPAs to be more or less normally distributed, if the tails are ignored.



3. The variable **rent** is best modeled by a Unif(400,900) random variable

   because it seems reasonable to expect any number (rounded to nearest dollar) with the range [$400,$900] might be reasonable for rent. An upper bound of $890 = MAX is totally reasonable too.



4. The variable **att** is best modeled by a Beta(7,3) random variable

   because ... the variable attention is measuring the *proportion of time* a student is paying attention. For estimating the parameters I knew that $1 < \beta < \alpha$ from the shape of the histogram. Then I played with the observation that mean(att) = 0.6722 and the mean of a Beta distribution was $\frac{\alpha}{\alpha+\beta} = .7$ with the parameters I tried above. This seemed quite close.