**Conceptual: Short answer questions. Be concise.**

1. How does knowing data type (e.g., qualitative vs. quantitative) provide insights into the type of distribution a random variable comes from?

**Answer:** For a qualitative (or categorical) data type, the probability distribution can be described by probability histogram, wherein the sum of all probabilities is equal to 1. For a quantitative (or continuous) data type, the probability distribution can be described by a density curve (e.g., uniform distribution and normal distribution curve) and the area under the curve is equal to 1.

1. What exactly does it mean to say that a data point is a "random variable is normal with a mean of 100 and standard deviation of 15"?

**Answer:** It means that the data point is independently drawn from an independent normal distribution where the mean is 100 and the standard deviation is 15.

**Applied: Show your code & plots**

The *central limit theorem* states that when independent random variables are added together, they sum to a normal distribution even if the original variables themselves are not normally distributed. For your homework test this assumption.

1. Using the *runif* function, run three experiments by simulating the outcomes of rolling a single, six-sided die. Show the distribution of each experiment. Show how the simulated means compare to the expected mean of a fair roll.
2. a) Exp 1: 10 independent throws
3. b) Exp 2: 1,000 independent throws

c) Exp 3: 10,000 independent throws

In [1]:

*# -------------------------------*

*# Question 3*

*# -------------------------------*

*# # Simulating rolling a single, six-sided die*

*a = 0*

*b = 6*

*ExpectedMean = (a+b)/2*

*# a) Exp 1: 10 independent throws*

*n = 10*

*dat1 = data.frame(length = ceiling(runif(n, min = a, max = b)))*

*ggplot(dat1, aes(length)) + geom\_density(alpha = 0.2) + ggtitle("10 times rolling") # Density curve for this experiment*

*SimulatedMean1 = mean(dat1$length) # simulated mean for experiment 1*

*ggplot(dat1, aes(length)) + geom\_histogram(alpha = 0.2) + stat\_bin(bins = 30) + ggtitle("10 times rolling") #Histogram*

*# b) Exp 2: 1,000 independent throws*

*n2 = 1000*

*dat2 = data.frame(length = ceiling(runif(n2, min = a, max = b)))*

*ggplot(dat2, aes(length)) + geom\_density(alpha = 0.2) + ggtitle("1000 times rolling") # Density curve*

*SimulatedMean2 = mean(dat2$length) # simulated mean for experiment 2*

*ggplot(dat2, aes(length)) + geom\_histogram(alpha = 0.2) + stat\_bin(bins = 30) + ggtitle("1000 times rolling")#Histogram*

*# c) Exp 3: 10,000 independent throws*

*n3 = 10000*

*dat3 = data.frame(length = ceiling(runif(n3, min = a, max = b)))*

*ggplot(dat3, aes(length)) + geom\_density(alpha = 0.2) + ggtitle("10000 times rolling") # Density curve*

*SimulatedMean3 = mean(dat3$length) # simulated mean for experiment 3*

*ggplot(dat3, aes(length)) + geom\_histogram(alpha = 0.2) + stat\_bin(bins = 30) + ggtitle("10000 times rolling") #Histogram*

1. Instead of rolling one die, run a set of experiments reporting the outcomes of rolling multiple dice at the same time.
2. a) Exp 1: 10,000 throws, 1 die
3. b) Exp 2: 10,000 throws, 2 dice
4. c) Exp 3: 10,000 throws, 3 dice

d) Exp 4; 10,000 throws, 6 dice.

Show the distribution of results for each experiment. Which of these experiments produces a distribution most similar to a normal normal distribution? Justify your conclusion using Q-Q plots.

**Answer:** The experiment 4 (throw 6 dice 10,000 times) produces a distribution most similar to a normal distribution. For each experiment, I ploted a Q-Q plot to visualize whether the data arise from a normal distribution. The Q-Q plot for experiment 4 looks most similar to a straight line, suggesting that its distribution is most similar to a normal distribution (see figures behind).

In [3]:

*# -------------------------------*

*# Question 4*

*# -------------------------------*

*# a = 0*

*b = 6*

*n = 10000*

*# Exp1: Throw 1 dice 10,000 times*

*Dice1 = data.frame(length = ceiling(runif(n, min = a, max = b))) # dice 1*

*dat1 = data.frame(length = Dice1$length)*

*ggplot(dat1, aes(length)) + geom\_density(alpha = 0.2) + ggtitle("Throw 1 dice 10000 times") # Density curve*

*ggplot(dat1, aes(length)) + geom\_histogram(alpha = 0.2) + stat\_bin(bins = 30) + ggtitle("Throw 1 dice 10000 times") #Histogram*

*ggplot(dat1, aes(sample=length)) + stat\_qq(distribution = qnorm) +*

*ggtitle("Throw 1 dice 10,000 times vs. normal sample") #QQ plot testing against a normal distribution*

*ggsave("Throw 1 dice.pdf", plot = last\_plot(), width = 6, height = 4)*

*# Exp2: Throw 2 dice 10,000 times*

*Dice1 = data.frame(length = ceiling(runif(n, min = a, max = b))) # dice 1*

*Dice2 = data.frame(length = ceiling(runif(n, min = a, max = b))) # dice 2*

*dat2 = data.frame(length = Dice1$length + Dice2$length)*

*ggplot(dat2, aes(length)) + geom\_density(alpha = 0.2) + ggtitle("Throw 2 dice 10000 times") # Density curve*

*ggplot(dat2, aes(length)) + geom\_histogram(alpha = 0.2) + stat\_bin(bins = 30) + ggtitle("Throw 2 dice 10000 times") #Histogram*

*ggplot(dat2, aes(sample=length)) + stat\_qq(distribution = qnorm) +*

*ggtitle("Throw 2 dice 10,000 times vs. normal sample") #testing against a normal distribution*

*ggsave("Throw 2 dice.pdf", plot = last\_plot(), width = 6, height = 4)*

*# Exp3: Throw 3 dice 10,000 times*

*Dice1 = data.frame(length = ceiling(runif(n, min = a, max = b))) # dice 1*

*Dice2 = data.frame(length = ceiling(runif(n, min = a, max = b))) # dice 2*

*Dice3 = data.frame(length = ceiling(runif(n, min = a, max = b))) # dice 3*

*dat3 = data.frame(length = Dice1$length + Dice2$length + Dice3$length)*

*ggplot(dat3, aes(length)) + geom\_density(alpha = 0.2) + ggtitle("Throw 3 dice 10000 times") # Density curve*

*ggplot(dat3, aes(length)) + geom\_histogram(alpha = 0.2) + stat\_bin(bins = 30) + ggtitle("Throw 3 dice 10000 times") #Histogram*

*ggplot(dat3, aes(sample=length)) + stat\_qq(distribution = qnorm) +*

*ggtitle("Throw 3 dice 10,000 times vs. normal sample") #testing against a normal distribution*

*ggsave("Throw 3 dice.pdf", plot = last\_plot(), width = 6, height = 4)*

*# Exp4: Throw 6 dice 10,000 times*

*Dice1 = data.frame(length = ceiling(runif(n, min = a, max = b))) # dice 1*

*Dice2 = data.frame(length = ceiling(runif(n, min = a, max = b))) # dice 2*

*Dice3 = data.frame(length = ceiling(runif(n, min = a, max = b))) # dice 3*

*Dice4 = data.frame(length = ceiling(runif(n, min = a, max = b))) # dice 4*

*Dice5 = data.frame(length = ceiling(runif(n, min = a, max = b))) # dice 5*

*Dice6 = data.frame(length = ceiling(runif(n, min = a, max = b))) # dice 6*

*dat4 = data.frame(length = Dice1$length + Dice2$length + Dice3$length + Dice4$length + Dice5$length + Dice6$length)*

*ggplot(dat4, aes(length)) + geom\_density(alpha = 0.2) + ggtitle("Throw 3 dice 10000 times") # Density curve*

*ggplot(dat4, aes(length)) + geom\_histogram(alpha = 0.2) + stat\_bin(bins = 30) + ggtitle("Throw 3 dice 10000 times") #Histogram*

*ggplot(dat4, aes(sample=length)) + stat\_qq(distribution = qnorm) +*

*ggtitle("Throw 6 dice 10,000 times vs. normal sample") #testing against a normal distribution*

*ggsave("Throw 6 dice.pdf", plot = last\_plot(), width = 6, height = 4)*







