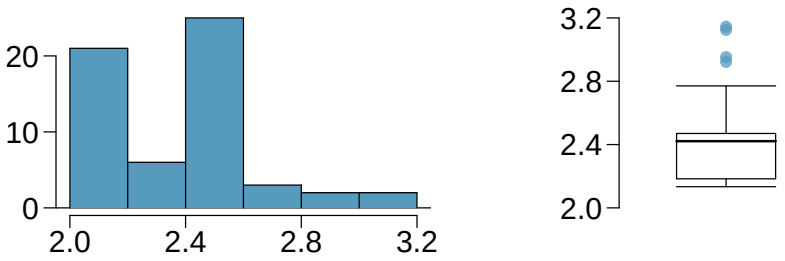
Sergio I Garcia Rios
Cornell University

Midterm 1

Name: _____

Write your responses in the spaces provided below. **WRITE LEGIBLY and SHOW ALL WORK!**

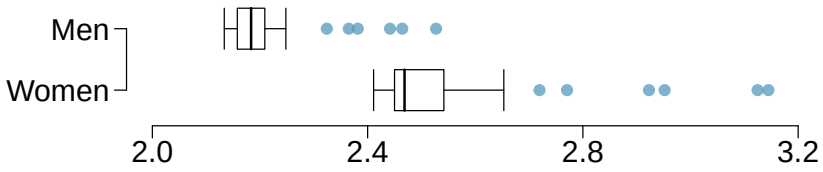
Marathon winners. The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1970 and 1999.



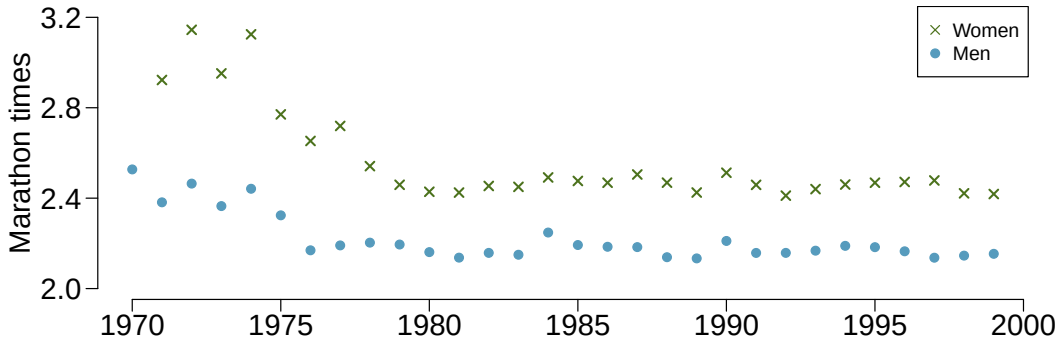
1. What features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?

2. What may be the reason for the bimodal distribution? Explain.

3. Compare the distribution of marathon times for men and women based on the box plot shown below.



4. The time series plot shown below is another way to look at these data. Describe what is visible in this plot but not in the others.



Fisher's irises. Sir Ronald Aylmer Fisher was an English statistician, evolutionary biologist, and geneticist who worked on a data set that contained sepal length and width, and petal length and width from three species of iris flowers (*setosa*, *versicolor* and *virginica*). There were 50 flowers from each species in the data set.

1. How many cases were included in the data?

2. How many numerical variables are included in the data? Indicate what they are, and if they are continuous or discrete.

3. How many categorical variables are included in the data, and what are they? List the corresponding levels (categories).

Coin flips. If you flip a fair coin 3 times, what is the probability of

1. getting all tails?
2. getting all heads?
3. getting at least one tails?

Swing voters. A 2012 Pew Research survey asked 2,373 randomly sampled registered voters their political affiliation (Republican, Democrat, or Independent) and whether or not they identify as swing voters. 35% of respondents identified as Independent, 23% identified as swing voters, and 11% identified as both.

1. Are being Independent and being a swing voter disjoint, i.e. mutually exclusive?

2. Draw a Venn diagram summarizing the variables and their associated probabilities.

3. What percent of voters are Independent but not swing voters?

4. What percent of voters are Independent or swing voters?

5. What percent of voters are neither Independent nor swing voters?

6. Is the event that someone is a swing voter independent of the event that someone is a political Independent?

Health coverage, frequencies. The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends. The following table summarizes two variables for the respondents: health status and health coverage, which describes whether each respondent had health insurance.

		Health Status					
		Excellent	Very good	Good	Fair	Poor	Total
Health Coverage	No	459	727	854	385	99	2,524
	Yes	4,198	6,245	4,821	1,634	578	17,476
	Total	4,657	6,972	5,675	2,019	677	20,000

1. If we draw one individual at random, what is the probability that the respondent has excellent health and doesn't have health coverage?

2. If we draw one individual at random, what is the probability that the respondent has excellent health or doesn't have health coverage?

Exit poll. Edison Research gathered exit poll results from several sources for the Wisconsin recall election of Scott Walker. They found that 53% of the respondents voted in favor of Scott Walker. Additionally, they estimated that of those who did vote in favor for Scott Walker, 37% had a college degree, while 44% of those who voted against Scott Walker had a college degree. Suppose we randomly sampled a person who participated in the exit poll and found that he had a college degree. What is the probability that he voted in favor of Scott Walker?

Male children. While it is often assumed that the probabilities of having a boy or a girl are the same, the actual probability of having a boy is slightly higher at 0.51. Suppose a couple plans to have 3 kids.

1. Use the binomial model to calculate the probability that two of them will be boys.

2. Write out all possible orderings of 3 children, 2 of whom are boys. Use these scenarios to calculate the same probability from part (1) but using the addition rule for disjoint outcomes. Confirm that your answers from parts (1) and (2) match.

Manufacturing workers. If 34% of NY manufacturing workers make more than \$19/hr, what is the probability that in a random sample of 100 NY manufacturing workers less than 30% make more than \$19/hr.

$p = 0.34$, $n = 100$
S/F: checks
$\mu = 0.34 \times 100 = 34$ *and* $\sigma = \sqrt{100 \times 0.34 \times 0.66} = 4.74$
$P\left(K < 30\right) = P\left(Z < \frac{30-34}{4.74}\right) = P(Z < -0.84) = 0.2$

Manufacturing workers. Government data indicates that the average hourly wage for manufacturing workers in the United States is \$18.61, with a standard deviation of \$1.35.

1. What Z score of manufacturing workers making more than \$20/hour?

   Given: $X_{US} \sim N(\mu = 18.61, \sigma = 1.35)$

   $$P(X > 20) = P\left(Z > \frac{20 - 18.61}{1.35}\right) = P(Z > 1.02) = 0.154 \rightarrow 15.4\%$$

2. Using the *Z-table* provided find the area under the bell curve to the left of z. *For instance: For a value of* $Z = 2.03$ *we would say that .9788 fall under the curve or 97.88% of the data.*

3. With that information, assess whether the amount of hourly worker making more than \$20 is high or low.

R functions. Looking at the following `R` function and given that $x = 17$

```
function(x){
  if(!is.atomic(x)){x <- x[,1]}
  y <- rep(0,length(x))
  y[x == "H"] <- 1
  y <- c(0, y, 0)
  wz <- which(y == 0)
  streak <- diff(wz) - 1
  return(data.frame(length = streak))
}
```

1. What would be the last item in the iteration?