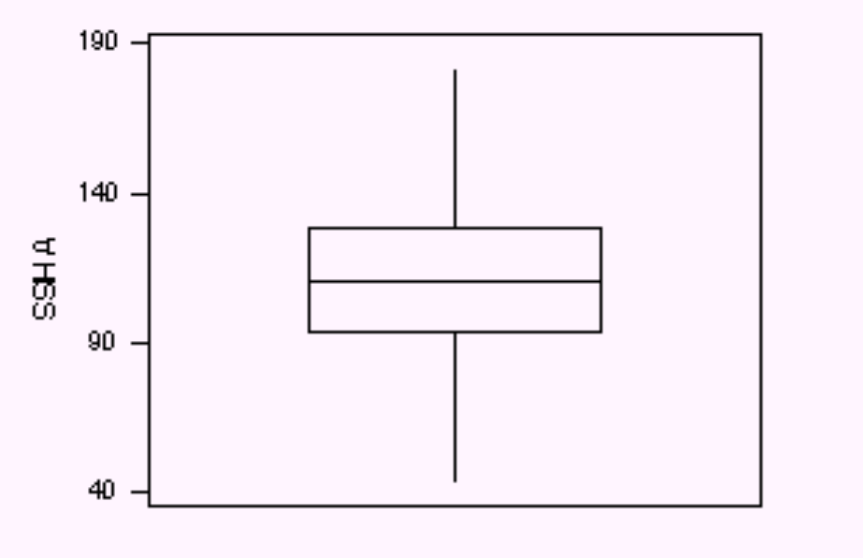
Solutions to Review Problems for Exam 1

1. The scores on the Survey of Study Habits and Attitudes (SSHA) for a sample of 150 first–year college women produced the following box plot



The statistical summary for the scores is

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
42.49	93.26	110.68	111.63	129.23	182.71

Estimate the number of women in the study with scores between 93.26 and 129.23.

> ***Answer:*** We want the number of data points between the first and third quartiles. This number corresponds to 50% of the data points, or 75 scores. Thus, 75 women in the study have scores between 93.26 and 129.23. □

2. The ages of the hourly paid workers at Westcaco involved in the second round of layoffs that the Envelope Division of the company went through in 1991 are listed here below in increasing order.

$$25, 33, 35, 38, 48, 55, \underline{55}, \underline{55}, 56, \underline{64}$$

The underlined numbers are the ages of the workers that were laid off in the second round.

(a) Compute the median age all the ages given above.

Answer: The median is the midpoint between the two middle ages in the ranked list; namely,

$$m = \frac{48 + 55}{2} = 51.5.$$

□

(b) Compute the median age the workers that were laid off and compared it to the median age of those that were kept.

Answer: The three workers that were laid off had ages 55, 55 and 64. Thus the median age of the laid off workers is 55
The ages of the workers that remained after the second round of layoffs are

$$25, 33, 35, 38, 48, 55, 56.$$

The median age for this group is 38. Thus, the median age for the workers that were laid off is much higher than that of the workers that remained. □

(c) Set up a procedure to test the hypothesis that Westvaco selected the three workers for layoffs at random using the median age as a test statistic.

Answer: Observe that the median age of the three laid off workers is 55.
Set up a replication procedure in which random samples of size 3 from the given set of ten ages are selected. We then compute the medians for each of the samples.
We then count the number of samples that yielded a median age of 55 or higher. □

(d) Define the p–value for the hypothesis test formulated in the previous part.

Answer: The p–value, in this case, is the probability that the median for a sample of size three drawn at random from the given ages is 55 or higher. □

(e) Describe a randomization procedure that you may use to estimate the p–value.

Answer: The p–value can be estimated by the proportion of times that the median of a randomly generated sample of size 3 from the 10 ages is 55 or higher. □

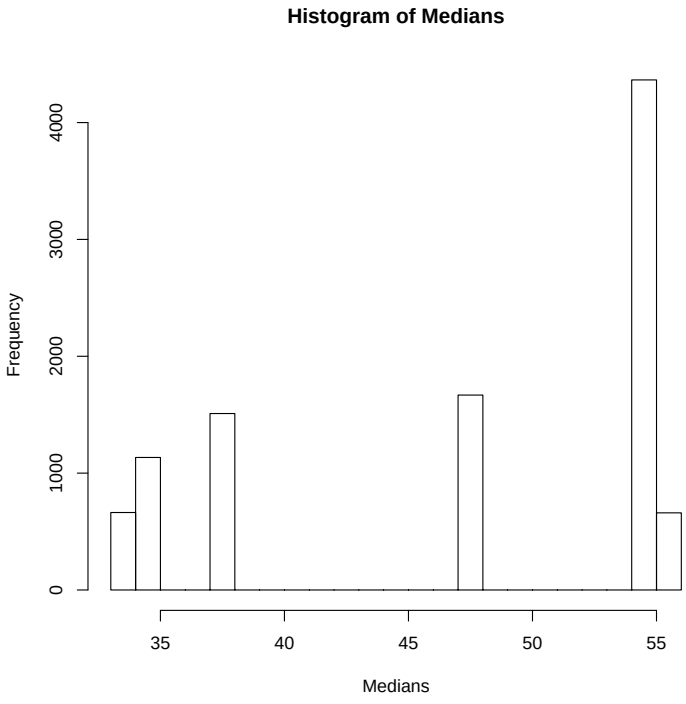


Figure 1: Sampling Distribution for the Median

(f) Figure 1 shows the histogram of the medians of the $10,000$ samples of size three selected at random from the 10 ages.

Use the histogram in Figure 1 to estimate the p–value. What do you conclude?

Solution: From the picture in Figure 1 we estimate that there are about 667 samples with a median above 55. This yields an estimate for the p value of about

$$p\text{–value} \approx \frac{667}{10000} = 0.0667$$

or about 6.67%. Thus, the p–value falls short of the 5% threshold for statistical significance. We therefore conclude that the data are not statistically significant at a 5% significance level is we use the median as a test statistic in a randomization test. Hence, the data involving median of the three hourly–paid workers involved the in the second round of layoffs do provide statistically evidence for age discrimination. □

3. Refer to the ages of the 10 hourly paid workers listed in the previous problem.

(a) Use the median of the ages of the workers that were laid off as a threshold to separate the workers into two classes: those whose age is above or equal to that value and those whose age is below the threshold. Based on this splitting, complete the Table 1

Age$\geqslant$ threshold? \ Fired?	No	Yes	Total
Yes	2		
No	5		
Total			

Table 1: Observed Values

Solution: Going through the 10 ranked ares given in Problem 2 and asking the question: "Is the age 55 or higher?" we find that the 10 can be grouped into two groups depending on whether the answer to the question was "yes" (Y), or "no" (N).

$$\text{N}, \text{N}, \text{N}, \text{N}, \text{N}, \text{Y}, \text{Y}, \text{Y}, \text{Y}, \text{Y}.$$

Age$\geqslant$ threshold? \ Fired?	No	Yes	Total
Yes	2	3	5
No	5	0	5
Total	7	3	10

Table 2: Observed Values

Out of the group labeled Y, 3 were laid off, while out of the group labeled N, none were laid off. We then obtain the values shown in Table 2.

□

(b) If the company did the selection at random, how many ages would you expect to see in each category in the table? Write your answers in Table 3

Age$\geqslant$ threshold? \ Fired?	No	Yes	Total
Yes			
No			
Total			

Table 3: Expected Values

Solution: Let X denote the number of workers in group Y that get selected for layoff in a random sample of size 3. The possible values for X are 0, 1, 2 and 3. To find the probability distribution for X, we compute

$$P(X=0) = \frac{\binom{5}{3}}{\binom{10}{3}} = \frac{1}{12};$$

$$P(X=1) = \frac{\binom{5}{1}\binom{5}{2}}{\binom{10}{3}} = \frac{5}{12};$$

$$P(X=2) = \frac{\binom{5}{2}\binom{5}{1}}{\binom{10}{3}} = \frac{5}{12};$$

$$P(X = 3) = \frac{\binom{5}{3}\binom{5}{0}}{\binom{10}{3}} = \frac{1}{12}.$$

We then obtain the probability distribution for X to be

$$P(X = k) = \begin{cases} 1/12 & \text{if } k = 0; \\ 5/12 & \text{if } k = 1; \\ 5/12 & \text{if } k = 2; \\ 1/12 & \text{if } k = 3. \end{cases} \tag{1}$$

The expected value for this random variable is

$$\begin{aligned} E(X) &= 0p_{_X}(0) + 1p_{_X}(1) + 2p_{_X}(2) + 3p_{_X}(3) \\ &= \frac{5}{12} + 2\frac{5}{12} + 3\frac{1}{12} \\ &= \frac{18}{12} = \frac{3}{2}, \end{aligned}$$

or 1.5. Thus, the entry in the "Yes" column and "Yes" row in Table 3 is 1.5. Similarly, the entry in the "Yes" column and "No" row in the table should be 1.5.

To find the entries in the "No" column of the table, we may proceed as in the previous part of the solution, or we may reason as follows: Seven out of the 10 workers get selected at random to keep their jobs. Since there are an equal number of workers of age 55 or above as there are workers below that age, there is a 1/2 chance for a worker selected to keep her or his job to be under the age of 55. Thus, on average, we expect $\frac{1}{2} \cdot 7$ workers selected to keep their jobs to be under 55. A similar reasoning leads to 3.5 workers selected to keep their jobs to be 55 or above. We then get the values shown in Table 4

□

4. Refer to the setup given in the previous problem.

   In the previous problem, you found that if you divide the hourly paid workers involved in the second round of layoffs into two groups: Older (age larger than or equal to a threshold) and Younger (age below the threshold); then, more

Age$\geqslant$ threshold? \ Fired?	No	Yes	Total
Older	3.5	1.5	5
Younger	3.5	1.5	5
Total	7	3	10

Table 4: Expected Values

people in the Older group were laid off than in the Younger group. Perform a test to determine whether this difference is statistically significant.

***Solution**:* Let the null hypothesis, H_o, be that the selection of the samples of size three is done at random. The test statistic is, X, the number of ages in the sample that are 55 or above. It then follows that the p–value is given by

$$p\text{–value} = P(X \geqslant 3) = P(X = 3) = \frac{1}{12} \approx 0.083,$$

or 8.3%. This falls short of the 5% threshold for statistical significance. Thus, the data for the 10 hourly–paid workers involved in the second round of layoffs at Westvaco provide evidence for age discrimination, but the evidence is not statistically significant. □

5. The table below shows a values of a random variable X and its probability distribution

Values of X	1	2	3
Probability	0.2	0.6	0.2

(a) Find the mean and standard deviation of X.

***Solution**:* The mean of the random variable X is given by

$$\mu = E(X) = 1 \cdot (0.2) + 2 \cdot (0.6) + 3 \cdot (0.2) = 2.0.$$

Therefore the variance of X is given by

$$\begin{aligned} \text{Var}(X) &= E[(X - \mu)^2] \\ &= (1 - 2)^2 \cdot (0.2) + (2 - 2)^2 \cdot (0.6) + (3 - 2)^2 \cdot (0.2) \\ &= 0.4. \end{aligned}$$

Thus, $\sigma_X = \sqrt{\text{Var}(X)} = \sqrt{0.4} \approx 0.63$. □

(b) Construct a different probability distribution with the same possible values, the same mean, and a larger standard deviation.

Solution: Consider the distribution for X given by □

Values of X	1	2	3
Probability	p	$1-2p$	p

for some value of p with $0 < p < 1$. The mean of X is still 2.0, by the symmetry of the distribution. The variance of the distribution is

$$\begin{aligned} \text{Var}(X) &= E[(X-\mu)^2] \\ &= (1-2)^2 \cdot p + (2-2)^2 \cdot (1-2p) + (3-2)^2 \cdot p \\ &= 2p. \end{aligned}$$

Thus, $\sigma_X = \sqrt{2p}$.

If we want σ_X to be larger than $\sqrt{0.4}$, we choose $p < 1$, so that

$$2p > 0.4 \quad \text{or} \quad p > 0.2.$$

(c) Construct a different probability distribution with the same possible values, the same mean, and a smaller standard deviation.

Solution: Based on the solution given in the previous part, for a smaller σ_X we choose $0 < p < 0.2$. □