

Homework 2

1. (*Binomial Test*) A random sample of tenth-grade boys resulted in the following 20 observed weights.

142	134	98	119	131
103	154	122	93	137
86	119	161	144	158
165	81	117	128	103

- a. Test the hypothesis that the median weight is 103;

Solution: $H_0: X_{.50} = 103$ vs. $H_a: X_{.50} \neq 103$

For the convenience of obtaining T_1 and T_2 , we first get the stem-and-leaf plot:

8	16
9	38
10	33
11	799
12	28
13	147
14	24
15	48
16	15

Accordingly, $T_1 =$ number of observations $\leq 103 = 6$

$T_2 =$ number of observations $< 103 = 4$

Since it is a two-tailed test, we have

$$\begin{aligned} \text{p-value} &= 2 \times \min \{ \Pr(Y \leq T_1), \Pr(Y \geq T_2) \} \\ &= 2 \times \min \{ \Pr(Y \leq 6), \Pr(Y \geq 2) \} \\ &= 2 \times \min \{ 0.05766, 0.9998 \} \\ &= 0.1153 > \alpha = 0.05, \end{aligned}$$

where $Y \sim \text{Binomial}(n = 20, p = 0.50)$. Thus we cannot reject the null at $\alpha = 0.05$.

- b. Apply the asymptotic Z test in part (a)

Solution: First, find $Z_1 = \frac{T_1 - np + 0.5}{\sqrt{np(1-p)}} = \frac{6 - 20 \times 0.5 + 0.5}{\sqrt{20 \times 0.5 \times (1-0.5)}} = -1.565$ and

$$Z_2 = \frac{T_2 - np + 0.5}{\sqrt{np(1-p)}} = \frac{4 - 20 \times 0.5 - 0.5}{\sqrt{20 \times 0.5 \times (1-0.5)}} = -2.9089. \text{ Hence we have}$$

$$\begin{aligned} \text{p-value} &= 2 \times \min \{ \Pr(Z \leq Z_1), \Pr(Y \geq Z_2) \} \\ &= 2 \times \min \{ \Pr(Z \leq -1.565), \Pr(Y \geq -2.9069) \} \\ &= 2 \times \min \{ 0.0588, 0.9982 \} \\ &= 0.1176 > \alpha = 0.05, \end{aligned}$$

where Z is the standard normal random variable. Thus we cannot reject the null at $\alpha = 0.05$.

- c. Test the hypothesis that the upper quartile is at least 150;

Solution: $H_0 : X_{.75} \geq 150$ vs. $H_a : X_{.75} < 150$

It can be found that $T_1 = \text{number of observations} \leq 150 = 16$

$T_2 = \text{number of observations} < 150 = 16$

With this one-sided hypothesis, it can be found

$$\text{p-value} = \Pr(Y \geq T_2) = 1 - \Pr(Y \leq 15) = 0.4148 > \alpha = 0.05,$$

where $Y \sim \text{Binomial}(n = 20, p = 0.75)$. Thus we cannot reject the null at $\alpha = 0.05$.

- d. Test the hypothesis that the third decile is no greater than 100.

Solution: $H_0 : X_{.30} \leq 100$ vs. $H_a : X_{.30} > 100$

It can be found that $T_1 = \text{number of observations} \leq 100 = 4$

$T_2 = \text{number of observations} < 100 = 4$

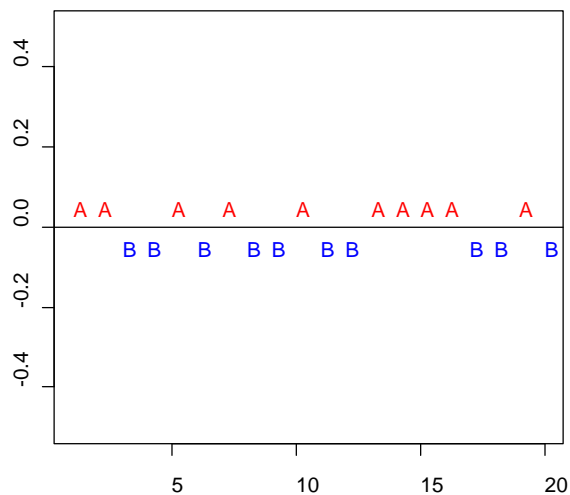
With this one-sided hypothesis, it can be found

$$\text{p-value} = \Pr(Y \leq T_1) = \Pr(Y \leq 4) = 0.2375 > \alpha = 0.05,$$

where $Y \sim \text{Binomial}(n = 20, p = 0.30)$. Thus we cannot reject the null at $\alpha = 0.05$.

- e. Apply the runs test to see whether the data seem to be a random sample.

Solution: It can be found that the sample median is 125. And the number of runs (in terms of above and below the median) is 12. Also, the parameters involved are: the total number of observations $N = 20$, number of observations Above $m = 10$, number of observations Below $n = 10$. Applying the asymptotic test, we have the standardized Runs test statistic $Z = 0.4595$ with p-value = 0.6459. Thus the data seem to be random.



2. Note that the *Binomial Test can be used to solve many other practical problems*. Here is another example. In a recent study examining colour preferences in infants, 30 babies were offered a choice between a red rattle and a green rattle. Twenty-five of the 30 selected the red rattle. Do these data provide evidence for a significant colour preference? Test at the 0.01 level of significance.

Solutions: *Step 1.* State the hypotheses, and specify alpha. The null hypothesis states that the proportion of babies preferring red rattles is not different from what is expected for a population where there is no preference for rattle colour. In symbols,

$$H_0 : p = p(\text{red}) = 0.5 \text{ and } q = p(\text{green}) = 0.5$$

The alternative hypothesis is that the proportions for the colour preferences are different from what is expected for these chance population proportions.

$$H_1 : p \neq 0.5 \text{ (and } q \neq 0.5)$$

We will set $\alpha = 0.01$.

Step 2. Locate the critical region. One heuristic way to determine if normal approximation could be used is to check if pn and qn are both greater than 10. Here, we can use the normal approximation to the binomial distribution. With $\alpha = 0.01$, the critical region is defined as any z-score value greater than +2.3263 or less than -2.3263.

Step 3. Calculate the test statistic. In the sample 25 out of 30 babies prefer the red rattle, so the sample proportion is

$$\frac{X}{n} = \frac{25}{30} = 0.83$$

The corresponding z-score is, (note that there is no need to apply the correction for discreteness in this problem.)

$$z = \frac{\frac{X}{n} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.83 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{30}}} = \frac{0.33}{0.09} = 3.66$$

Step 4. Make a decision about H_0 , and state a conclusion. The obtained z-score is in the critical region. Therefore, we reject the null hypothesis. On the basis of these data, we conclude that the proportion of babies preferring red rattles is significantly different from the proportion of babies preferring green rattles.

Twenty five out of 30 babies preferred red rattles. A binomial test revealed that there is a significant preference for red rattles since $z = 3.66$, $p < 0.01$.

Testing for Normality

How does one know if the data fits a normal distribution?

Cholesterol Data from the Framingham Heart Study

Stem-and-leaf plot	Freq	Cumul Freq
16 7	1	1
17	0	1
18 4	1	2
19 28	2	4
20 02	2	6
21 0125678	7	13
22 0556	4	17
23 0000122244668	13	30
24 03678	5	35
25 444668	6	41
26 347778	6	47
27 00288	5	52
28 35	2	54
29	0	54
30 008	3	57
31	0	57
32 7	1	58
33 46	2	60
34	0	60
35 3	1	61
36	0	61
37	0	61
38	0	61
39 3	1	62

Descriptive Statistics:

$$n = 62, \bar{X} = 250.03, s = 41.44, \sqrt{b_1} = \frac{m_3}{m_2^{3/2}} = 1.024, b_2 = \frac{m_4}{m_2^2} = 4.577 \text{ where } m_k = \frac{\sum (X_i - \bar{X})^k}{n}$$

Tests to Consider

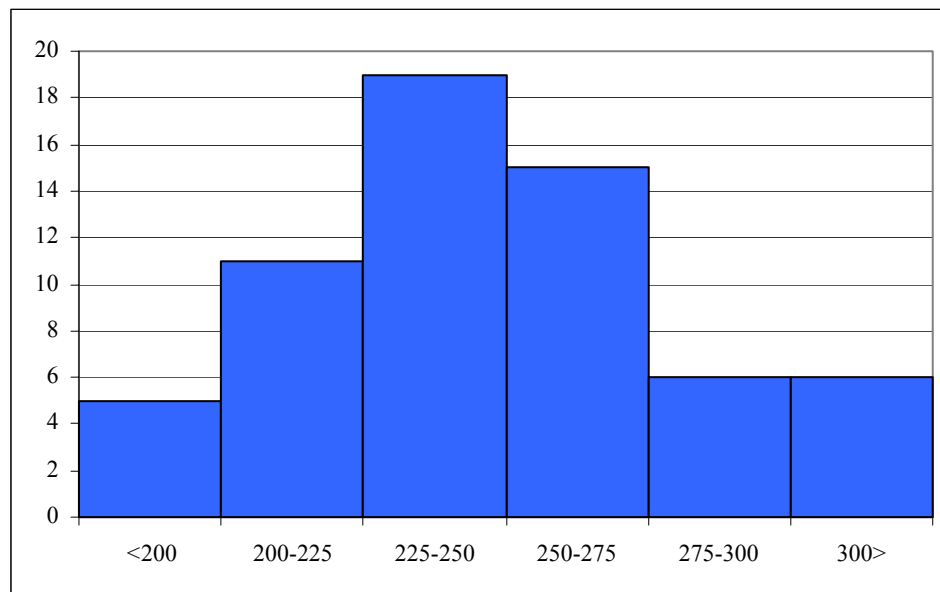
- [1] Chi-Square Goodness-of-Fit Test
- [2] Graphics – Normal Probability Plot
- [3] Non-Parametric Kolmogorov-Smirnov Test
- [4] Geary's Test

[1] Chi-Square Goodness-of-Fit Test

mean 250.0323
 stdev 41.44321

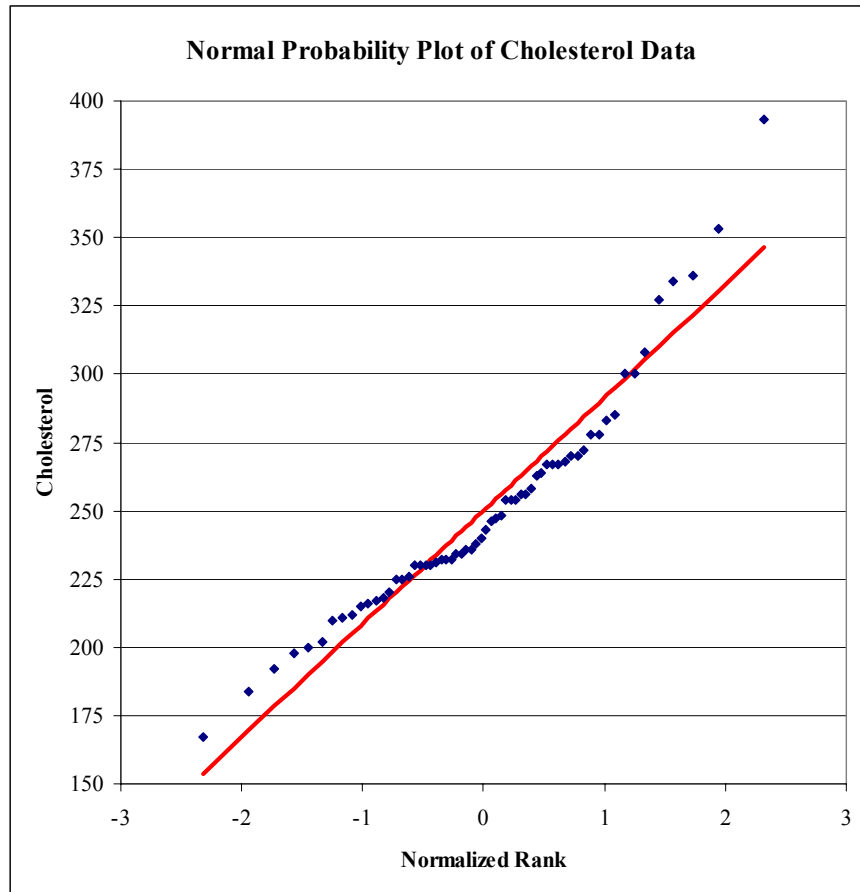
Bin	Obs Freq	z	Exp Freq	$(O_i - E_i)^2 / E_i$
<200	5	-1.20725	7.04743	0.594822
200-225	11	-0.60401	9.873446	0.128539
225-250	19	-0.00078	14.05987	1.735782
250-275	15	0.602457	14.06628	0.061981
275-300	6	1.205692	9.886948	1.528112
300>	6		7.066027	0.160828
sum	62		62	$\chi^2 = 4.210064$ p = 0.239656

For example $-1.2075 = (200 - \text{mean})/\text{stdev}$ and $7.04743 = 62 * \text{pnorm}(-1.20725, 0, 1)$



By this test the data would be accepted as fitting a normal distribution.
 Note that the values in the tails had to be combined to assure at least 5 in each cell.

[2] Normal Probability Plot

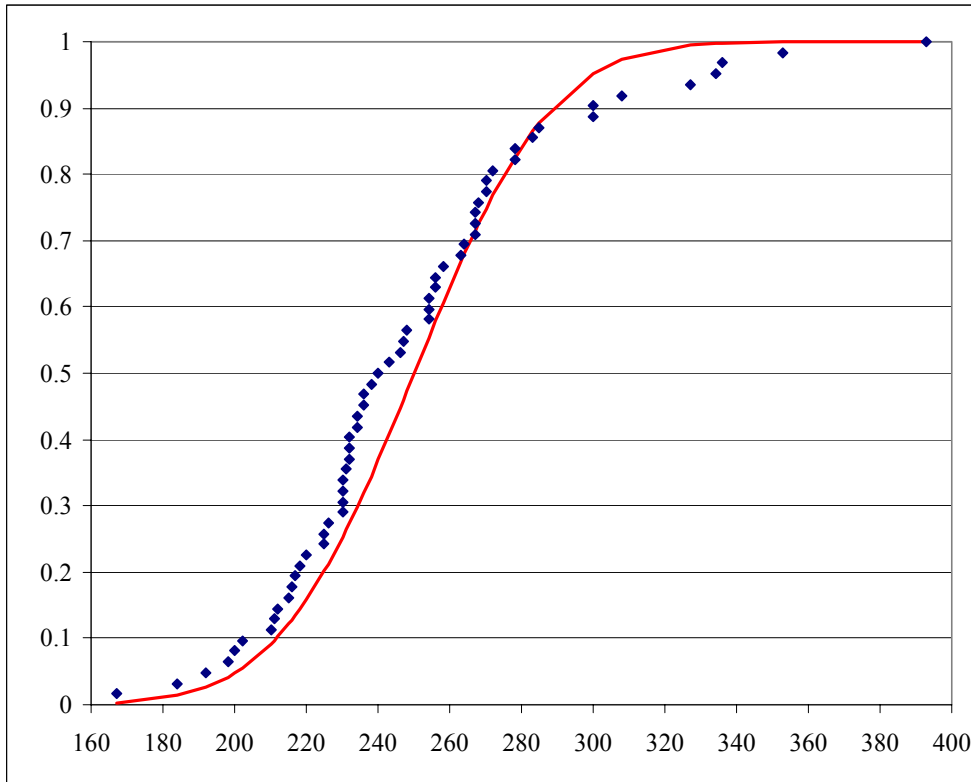


i	$(i-0.375)/(n+0.25)$	$Z_i = \Phi^{-1}[(i-0.375)/(n+0.25)]$	Data	$m + Z_i s$
1	0.010040	-2.324844	167	153.6833
2	0.026104	-1.941408	184	169.5741
3	0.042169	-1.726056	192	178.4990
...
61	0.973896	1.941408	353	330.4904
62	0.989960	2.324844	393	346.3812

The normal probability plot is a plot of $Z = \Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right)$ on $X_{(i)}$ where $X_{(i)}$ is the i^{th} ordered sample $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ and Z is the value such that $\frac{i-3/8}{n+1/4} = \int_{-\infty}^Z \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ for $I=1, 2, \dots, n$. In the above plot the skewness to the right is very evident and indicates that the data set may not be normally distributed.

It also can be found (using R) that the Shapiro-Wilk test is $W=.9386$ with p -value 0.003898 .

[3] Non-parametric Kolmogorov-Smirnov Test



Data	cumul f	normal	D+	D-
167	0.016129	0.001517	-0.0146	0.0015
184	0.032258	0.009208	-0.0230	-0.0069
192	0.048387	0.019159	-0.0292	-0.0131
...				
236	0.467742	0.30853	-0.1592	-0.1269
...				
353	0.983871	0.999883	-0.0158	0.0321
393	1	0.999999	-0.0000	0.0161

$$D_n^+ = \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - F(x_{(i)}, \theta) \right\}, D_n^- = \max_{1 \leq i \leq n} \left\{ F(x_{(i)}, \theta) - \frac{i-1}{n} \right\}, D_n = \max(D_n^+, D_n^-)$$

$$P\{\sqrt{n}D_n > t\} \rightarrow K(t) = 2\left(e^{-2t^2} - e^{-8t^2} + e^{-18t^2} - \dots \pm e^{-2k^2t^2} \pm \dots\right)$$

a fully specified normal distribution

Using a normal distribution with an arbitrary parameters $\mu = 250$ and $\sigma = 28$ gives the graph above, a value of $D_n = 0.1592$ for $n = 62$ and a p-value of 0.0863. This is only a marginal indication of non-normality. This test may be used for any probability distribution. This test requires presumed values of μ and σ and should not be based upon the sample data. Other tests are more powerful because they are for made for a specific distribution.