MATH 308 Assignment 2:
R Simulations

Nakul Joshi

January 23, 2014

Source code from file `assignment2.R`.
Result: The probability of obtaining a run of 7 heads in a sequence of 100 coin flips is ≈ 0.32.

Math 308 Assignment 4
Exercises 2.8

Nakul Joshi

February 17, 2014

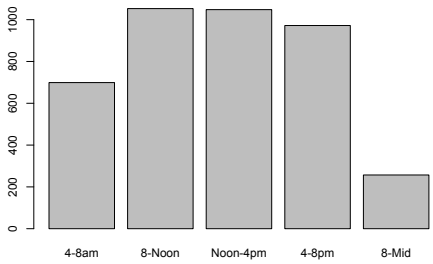
2

$\overline{x} = 6.5$
$m = 5.5$
$\tilde{x} = 2.389726$
$\tilde{m} = 2.342779$
$f(\overline{x}) \neq \tilde{x}$
$f(m) \neq \tilde{m}$

4

a)

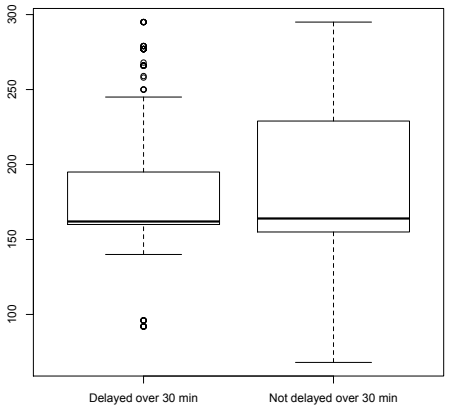
4-8am	8-Noon	Noon-4pm	4-8pm	8-Mid
699	1053	1048	972	257



b)

	No	Yes	Proportion
Mon	569	61	0.09682540
Tue	535	93	0.14808917
Wed	488	76	0.13475177
Thu	434	132	0.23321555
Fri	493	144	0.22605965
Sat	406	47	0.10375276
Sun	507	44	0.07985481

c)



d)

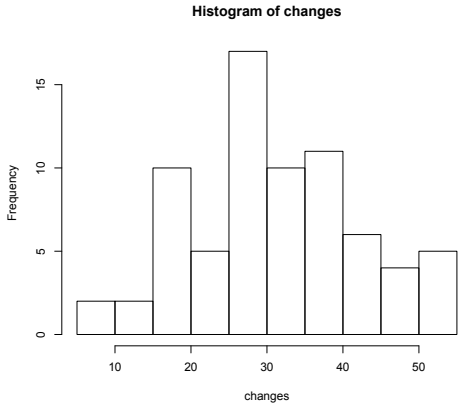
There appears to be no relationship.

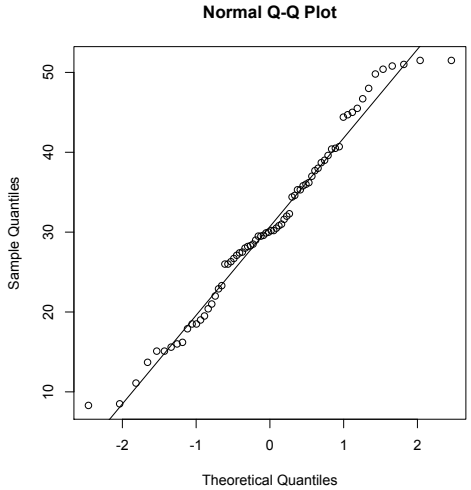
6

a)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8.30	23.20	30.10	30.93	38.17	51.50

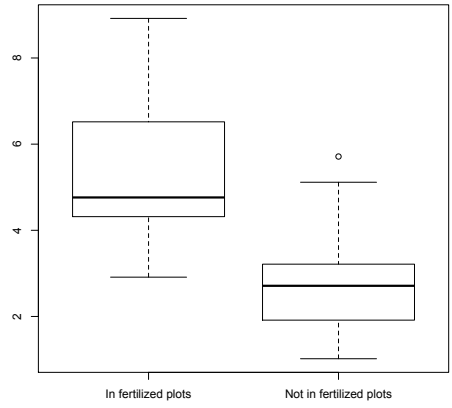
b)





The distribution is approximately normal as shown by the close fit between the normal and theoretical quantiles.

c)



d)

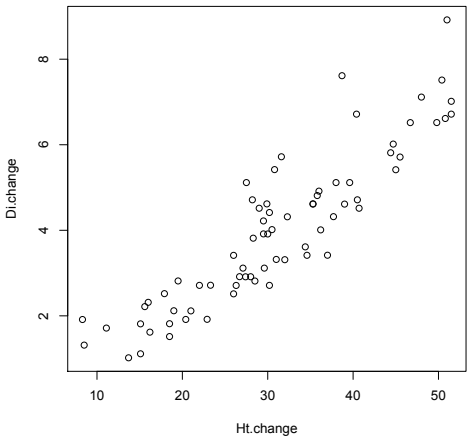
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.912	4.318	4.762	5.274	6.518	8.919

Table 1: Summary of F

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.019	1.915	2.712	2.718	3.165	5.712

Table 2: Summary of NF

e)



The diameter changes roughly increase with the height changes.

8

a)

To find the median, we need a value m such that, for $a = 1/2$:

$$\begin{aligned} a &= \int_{-\infty}^{m} f(x)\,\mathrm{d}x \\ &= \int_{0}^{m} \lambda e^{-\lambda x}\,\mathrm{d}x \\ &= 1 - e^{-\lambda m} \\ \implies e^{-\lambda m} &= 1 - a \\ \implies -\lambda m &= \log(1-a) \\ \implies m &= \lambda^{-1} \log \frac{1}{1-a} \\ &= \lambda^{-1} \log 2 \end{aligned}$$

Similarly, for first and third quartiles, we use $a = 1/4$, and $a = 3/4$, respectively:

$$\begin{aligned} Q_1 &= \lambda^{-1} \log \frac{1}{1-\frac{1}{4}} = \lambda^{-1} \log \frac{4}{3} \\ Q_3 &= \lambda^{-1} \log \frac{1}{1-\frac{3}{4}} = \lambda^{-1} \log 4 \end{aligned}$$

b)

As with the previous problem:

$$\begin{aligned} a &= \int_{-\infty}^{m} f(x)\,\mathrm{d}x \\ &= \int_{1}^{m} \frac{\alpha}{x^{\alpha+1}}\,\mathrm{d}x \\ &= 1 - m^{-\alpha} \\ \implies m &= \sqrt[\alpha]{\frac{1}{1-a}} \\ &= \sqrt[\alpha]{2} \\ Q_1 &= \sqrt[\alpha]{\frac{4}{3}} \\ Q_3 &= \sqrt[\alpha]{4} \end{aligned}$$

13

The pmf is $f(x) = \binom{20}{x} 0.3^x 0.7^{20-x}$. So, the cdf $F(x) = \sum_{i=0}^{x} f(i)$. Then, there is no such q, since 2 is too small ($F(2) < 0.04$) and 3 is too large ($F(3) > 0.1$).

14

a)

One the QQ plot, the points showed a close fit to the straight line, but the histogram was never symmetric, only sometimes unimodal, and rarely mound-shaped.

b)

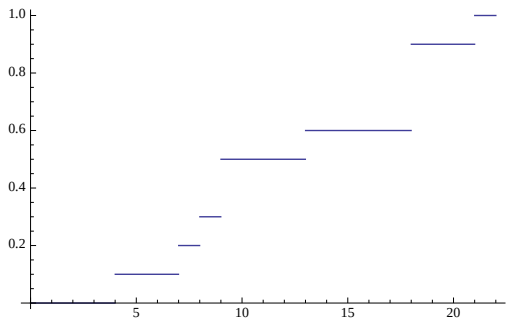
While the quality of the fit stayed good, the histograms only had a marginal improvement in terms of resembling a normal distribution (symmetric, unimodal and mound-shaped).

c)

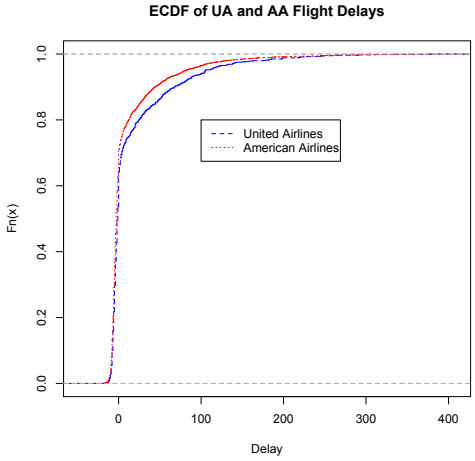
It seems that QQ-normal plots are better for visually determining whether or not data is normally distributed.

15

x	f	cf	cdf=cf/total
4	1	1	0.1
7	1	2	0.2
8	1	3	0.3
9	2	5	0.5
13	1	6	0.6
18	3	9	0.9
21	1	10	1.0



17



From the plot, we can see that for a small range of delay times, the ecdf is higher for AA than for UA.

Math 308 Assignment 5
Salk Vaccine Trial Hypothesis Testing

Nakul Joshi

16th February 2014

1 Hypergeometric Distribution

Under the null hypothesis, we can assume that everyone who developed the disease after the trial would have done so regardless of which group they were put into. Thus, the number of cases x in the treatment group follows a hypergeometric distribution where:

Population size $N = 200000 + 200000 = 400000$

Cases in population $K = 56 + 141$

Size of treatment group $n = 200000$

Then, the pmf is:

$$Pr_H(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

The requisite p-value is then the probability of 56 or fewer of the cases being chosen into the control group:

$$\begin{aligned} p_H &= \sum_{i=0}^{56} Pr_H(X = i) \\ &= 5.98 \times 10^{-10*} \end{aligned}$$

This value is certainly statistically significant, and so we reject the null hypothesis that the Salk vaccine is ineffective.

2 Binomial Approximation

Even though the trial involves sampling with replacements, the large sample size means that, under the null, the the number of cases in the treatment group can be modelled by the binomial distribution:

Proportion of cases $p = K/N = \frac{197}{400000}$

The pmf is then:

$$Pr_B(X = k) = \binom{n}{k} p^k (1-p)^{K-k}$$

The p-value is:

$$\begin{aligned} p_B &= \sum_{i=0}^{56} Pr_N(X = i) \\ &= 2.26 \times 10^{-6} \end{aligned}$$

This is four orders higher than p_H, but is still extremely low.

3 Normal Approximation

We can further approximate x with normally distributed variable y having $\mu = \frac{nK}{N}$ and $\sigma = \frac{\sqrt{nK(N-K)}}{N}$, giving the pdf:

$$Pr_N(Y = y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Applying the continuity correction:

$$\begin{aligned} p_N &= \int_0^{56+0.5} Pr_N(y)\,\mathrm{d}y \\ &= 1.00 \times 10^{-15} \end{aligned}$$

Or five orders lower than p_H.

*This result was obtained via the *Mathematica* software package.

Math 308 Assignment 6
Smoking/Birth-weight Correlation Testing

Nakul Joshi

16th February 2014

Running the permutation test gives us a p value of 0, which allows us to reject the null hypothesis that Tobacco use by the mother does not affect the birth weight of newborns.

Math 308 Assignment 7
Exercises 3.9

Nakul Joshi

February 17, 2014

4

The null hypothesis is that the difference in proportions is zero. However, performing the permutation test gave a p-value of 0.002, allowing us to reject the null at 1% confidence. Thus, the difference in proportions is statistically significant.

8

The p-value is 1, which does not let us reject the null hypothesis that the presence of competition has no value on the height change of the seedlings.

11

Null Hypothesis Voting preference is independent of age.

Alternative hypothesis Voting preference depends on age.

Age	Response		
	For	Against	All
18-29	172	52	224
30-49	313	103	416
50+	258	119	377
All	743	274	1017

Table 1: Observed values

Multiplying column marginal fractions by row marginal totals, we can get the expected values:

Age	Response	
	For	Against
18-29	164	60
30-49	304	112
50+	275	102

Table 2: Expected values

Then, we calculate the χ^2 test statistic:
$c = \sum_{i,j}^{\text{all cells}} \frac{(\text{observed}_{i,j} - \text{expected}_{i,j})^2}{\text{expected}_{i,j}} = 6.33$

Under the null, C follows a χ^2 distribution with $(3-1)\times(2-1) = 2$ degrees of freedom; i.e. $C \sim \chi^2_2$. So, the p-value is $P(C > c) = \int_c^\infty \frac{e^{-t/2}}{2}\,\mathrm{d}t \approx 0.042$.

Thus, we can reject the null at 5% significance, but not at 1% significance.

13

a)

We are testing for homogenity since we want to know whether the distribution of fin ray counts differs from lake to lake.

b)

Null hypothesis Fin ray distributions are the same from lake to lake.

Alternative hypothesis Fin ray distributions are different from lake to lake.

Habitat	Ray Count						
	36	35	34	33	32	31	All
Guadalupe	14	30	42	78	33	14	211
Cedro	11	28	53	66	27	9	194
San Clemente	10	17	61	53	22	10	173
All	71	110	190	230	114	64	779

Habitat	Ray Count					
	≥36	35	34	33	32	≤31
Guadalupe	19	30	51	62	31	17
Cedro	18	27	47	57	28	16
San Clemente	16	24	42	51	25	14

Table 3: Expected Values

$c = \sum_{i,j}^{\text{all cells}} \frac{(\text{observed}_{i,j} - \text{expected}_{i,j})^2}{\text{expected}_{i,j}} = 41.77$, where $C \sim \chi^2_{10}$. So, $p = P(C > c) = \int_c^\infty \frac{t^{10/2-1}e^{-t/2}}{2^{10/2}\Gamma(10/2)}\,\mathrm{d}t = \int_c^\infty \frac{t^4 e^{-t/2}}{768}\,\mathrm{d}t = 8 \times 10^{-6}$. So, we can reject the null.

17

a)

Happiness	Gender	
	Female	Male
Not too happy	109	61
Pretty happy	406	378
Very happy	205	210

Table 4: Observed happiness against gender data

b)

We get a p value of 0.004, allowing us to reject the null hypothesis of independence at a 1% significance level.

19

a)

Let the elements of the contingency table be $o_{i,j}$, and let their corresponding row and column totals be r_i and c_j respectively. Further, let the total number of observations $\sum_{i,j} o_{i,j} = n$. The corresponding expected values are then $e_{i,j} = \frac{r_i c_j}{n}$, which gives the test statistic $c = \sum_{i,j} \frac{(o_{i,j}-e_{i,j})^2}{e_{i,j}}$

However, if each element is multiplied by k, then $o_{i,j}$,r_i, c_j and n are each multiplied by k. So, the corresponding expected values become $e^*_{i,j} = \frac{k^2}{k} \times \frac{r_i c_j}{n} = k \times e_{i,j}$. The new test statistic $c^* = \sum_{i,j} \frac{(o^*_{i,j}-e^*_{i,j})^2}{e^*_{i,j}} = \sum_{i,j} \frac{k^2}{k} \times \frac{(o_{i,j}-e_{i,j})^2}{e_{i,j}} = k \times c$. Thus, the test statistic is also multiplied by k.

However, the marginal probabilities are unchanged since the k's cancel on the row and overall totals. Further, the degrees of freedom are unaffected since they only depend upon r and c.

b)

Originally, the p-value was $\int_c^\infty f(t;k)\,\mathrm{d}t$, but the new p-value becomes $\int_{kc}^\infty f(t;k)\,\mathrm{d}t$. Thus, the p value reduces, since $k > 1 \implies kc > c$.

22

a)

p	q
0.2	16.1
0.4	20.2
0.6	23.8
0.8	27.9

b)

Interval	Counts	
	Obs.	Exp.
<16.11	16	10
16.11–20.23	13	10
20.23–23.77	9	10
23.77–27.89	9	10
>27.89	3	10

c)

We get a p-value of 0.048, which does not let us reject the hypothesis that the data was drawn from a $N(22, 7^2)$ distribution.

25

We get a p-value of 0.78, which does not let us reject the hypothesis that the numbers are uniformly distributed.

29

a)

Let N_{ij}be the element at row i and column j. Then, since there are only two rows and two columns, let the remaining row and column indices be $\bar{\imath}$ and $\bar{\jmath}$ respectively. Then, the row total is $R_i = N_{ij} + N_{i\bar{\jmath}}$ and the column total is $C_j = N_{ij} + N_{\bar{\imath}j}$. So, the expected value at i, j is

$$\begin{aligned}\breve{E}[N_{ij}] &= \frac{R_i \times C_j}{n} \\ &= \frac{(N_{ij} + N_{i\bar{\jmath}})(N_{ij} + N_{\bar{\imath}j})}{n}\end{aligned}$$

This gives:

$$\begin{aligned}&N_{ij} - \breve{E}[N_{ij}] \\ &= N_{ij} - \frac{(N_{ij} + N_{i\bar{\jmath}})(N_{ij} + N_{\bar{\imath}j})}{n} \\ &= \frac{N_{ij}(\cancel{N_{ij}} + \cancel{N_{i\bar{\jmath}}} + \cancel{N_{ij}} + N_{\bar{\imath}\bar{\jmath}}) - (\cancel{N_{ij}^2} + \cancel{N_{ij}N_{i\bar{\jmath}}} + \cancel{N_{ij}N_{i\bar{\jmath}}} + N_{i\bar{\jmath}}N_{\bar{\imath}j})}{n} \\ &= \frac{N_{ij}N_{\bar{\imath}\bar{\jmath}} - N_{i\bar{\jmath}}N_{\bar{\imath}j}}{n}\end{aligned}$$

Thus:

$$(N_{ij} - \breve{E}[N_{ij}])^2 = \left(\frac{N_{ij}N_{\bar{\imath}\bar{\jmath}} - N_{i\bar{\jmath}}N_{\bar{\imath}j}}{n}\right)^2$$

This expression is symmetrical; i.e., it does not change by swapping i and $\bar{\imath}$ or j and $\bar{\jmath}$. So, it is equivalent to:

$$\left(\frac{N_{11}N_{22} - N_{21}N_{12}}{n}\right)^2$$

which is independent of i, j $\square$.

b)

Call the previously obtained expression k. Then,

$$\begin{aligned}C &= kn/R_1C_1 + kn/R_1C_2 + kn/R_2C_1 + kn/R_2C_2 \\ &= kn\frac{R_1C_1 + R_1C_2 + R_2C_1 + R_2C_2}{R_1R_2C_1C_2} \\ &= kn\frac{R_1(C_1 + C_2) + R_2(C_1 + C_2)}{R_1R_2C_1C_2} \\ &= kn\frac{n(R_1 + R_2)}{R_1R_2C_1C_2} = k\frac{n^3}{R_1R_2C_1C_2} \\ &= \frac{(N_{11}N_{22} - N_{21}N_{12})^2}{n^2}\frac{n^3}{R_1R_2C_1C_2} \\ &= n(N_{11}N_{22} - N_{21}N_{12})^2/R_1R_2C_1C_2 \quad \square\end{aligned}$$

c)

Via R, the expression is verified since both methods yield $C \approx 0.0234$