MATH 308 Assignment 12: Sequences in Coin Flips

Nakul Joshi

March 26, 2014

1 Expected Number of Occurences

For each coin flip i, let X_i be the event that the sequence 01111111 occurs starting at position i. Then:

$$\mathbb{E}\left(\mathbb{1}(X_i)\right) = \begin{cases} 2^{-8}, & \text{if } 1 \leq i \leq 93 \\ 0, & \text{if } i > 93 \end{cases}$$

So, the expected number of occurences is:

$$\begin{aligned} N &= \mathbb{E}\left(\sum_{i=1}^{100} \mathbb{1}(X_i)\right) \\ &= \sum_{i=1}^{100} \mathbb{E}\left(\mathbb{1}(X_i)\right) \\ &= \sum_{i=1}^{93} \mathbb{E}\left(\mathbb{1}(X_i)\right) \\ &= 93/2^8 \approx 0.363 \end{aligned}$$

2 Probability of Occurence

Using the result above, we can approximate the coin-flipping experiment as a Poisson process with rate parameter $\lambda = 93/2^8$.

The requisite probability can then be calculated as `1-ppois(0,93/2^8)`, which comes out to ≈ 0.30, which is slightly under the result from Assignment 2.

MATH 308 Assignment 13: Comparing Estimator Biases

Nakul Joshi

March 26, 2014

The bias when dividing by $n-1$ was found to be 0.003, while the bias when dividing by n was found to be -0.046, which is two orders higher. Thus, dividing by $n-1$ is preferred for reducing bias.

MATH 308 Assignment 14: Distribution of Second Largest Element

Nakul Joshi

March 26, 2014

1 Distribution of Second-Largest Element

Let the distribution being sampled have pdf $f(x)$, and let $X = \{X_1, X_2, X_3 \ldots X_n\}$ be the set of random variables drawn from that distribution. Further, let Y be the second largest element of X. Then, for any value x, we have:

$$
\begin{aligned}
P(Y = x) &= P(\text{some } X = x \;\&\; n-2 \text{ of other } X < x) \\
&= \sum_{i=1}^{n} P(X_i = x) P(n-2 \text{ of other } X < x)) \\
&= nP(X_1 = x) \\
&\binom{n-1}{n-2} P^{n-2}(X_1 < x) P(X_1 > x) \\
&= \frac{n(n-1)!}{1!(n-2)!} f(x) F^{n-2}(x)(1 - F(x)) \\
\implies f_Y(x) &= n(n-1) f(x) F^{n-2}(x)(1 - F(x))
\end{aligned}
$$

2 Expected value from $\mathcal{U}(1, 2)$

We know that $f(x) = 1$ and $F(x) = x$ in the range $[0, 1]$. Therefore,

$$
\begin{aligned}
f_Y(x) &= n(n-1) f(x) F^{n-2}(x)(1 - F(x)) \\
&= n(n-1)(1)(x^{n-2})(1 - x) \\
&= n(n-1)x^{n-1}(1 - x)
\end{aligned}
$$

$$
\begin{aligned}
\implies \mathbb{E}(Y) &= \int_0^1 n(n-1)x^{n-2}(1 - x)\,\mathrm{d}x \\
&= n(n-1) \int_0^1 x^{n-1} - x^n \,\mathrm{d}x \\
&= n(n-1) \left(\frac{1}{n} - \frac{1}{n+1} \right) \\
&= \cancel{n}(n-1) \frac{\cancel{n+1-n}}{\cancel{n}(n+1)} \\
&= \frac{n-1}{n+1}
\end{aligned}
$$

MATH 308 Assignment 15:
Exercises 5.10

Nakul Joshi

March 26, 2014

5

a)

Suppose that cookie orders are represented such that, for example, an order of 1 sugar, 2 chocolate chips, 0 oatmeal, 0 peanut butter, and 2 ginger snaps is given by $|c|cc|cc|$. Then, two of the bars are fixed, giving the string $|???????|$, where the nine ? symbols are to be replaced by five c's and two $|$'s. If we choose 5 distinct positions in which to place the c's, the positions of the remaining $|$'s is fixed. Thus, the number of possible cookie orders is $\binom{9}{5}$.

b)

More generally, we can use 'cookie types' to represent occurances of particular elements in the chosen sets. Then, fixing the position of two $|$'s leaves $n+1-2 = n-1$ $|$'s and n c's, for a total of $2n-1$ symbols. Picking element members requires n position choices, giving the total number of set choices as $\binom{2n-1}{n}$.

c)

For a sample S of size n, each bootstrap sample represents the choice of n elements from S with replacement. Thus, the number of bootstrap samples is as above $\binom{2n-1}{n}$.

6

a)

The number of bootstrap samples with k_1 a_1's, k_2 a_2's $\ldots k_n$ a_n's is the same as the number of permutations of the string

$$\underbrace{\overbrace{a_1a_1\ldots a_1}^{k_1\text{times}}\overbrace{a_2a_2\ldots a_2}^{k_2\text{times}}\ldots\overbrace{a_na_n\ldots a_n}^{k_n\text{times}}}_{n\text{ symbols}}$$

Using the formula for string permuations with repeated characters, this yields

$$\frac{n!}{k_1!k_2!\ldots k_n!}$$

which, by definition, equals

$$\binom{n}{k_1, k_2 \ldots k_n} \quad \square$$

b)

A bootstrap sample with k_i a_i's requires choosing k_i of n positions, after which the remaining $n-k_i$ positions can be filled by any of the remaining $n-1$ elements. Thus, the number of such samples is:

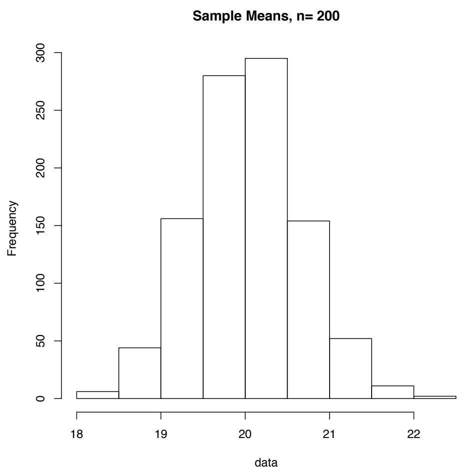
$$(n-1)^{n-k_i}\binom{n}{k_i}$$

The total number of bootstrap samples is simply n^n. Thus, the required probability is

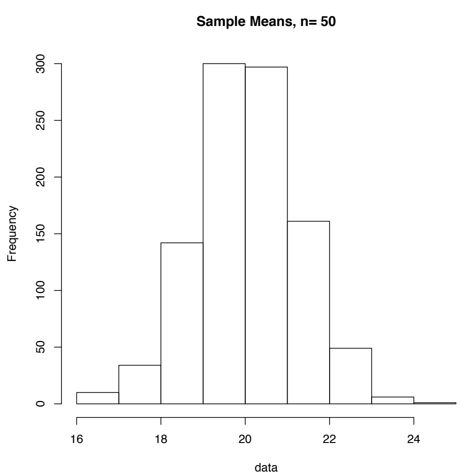
$$\frac{(n-1)^{n-k_i}}{n^n}\binom{n}{k_i}$$

8

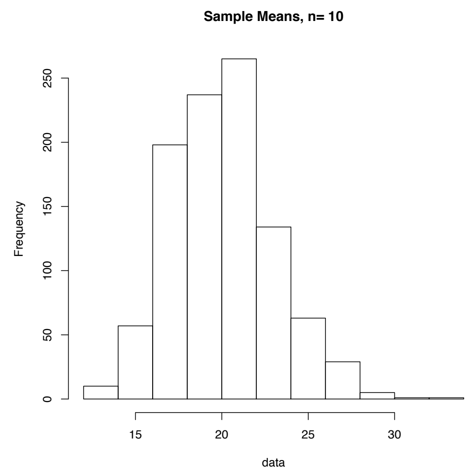
a)



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18.21	19.58	20.02	20.02	20.44	22.49



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
16.19	19.26	20.03	20.03	20.87	24.43



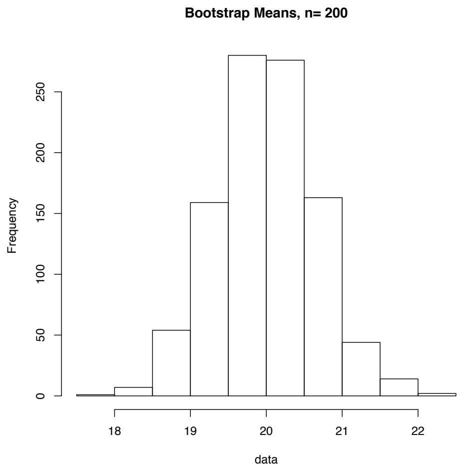
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
12.56	17.87	19.99	20.05	21.89	33.20

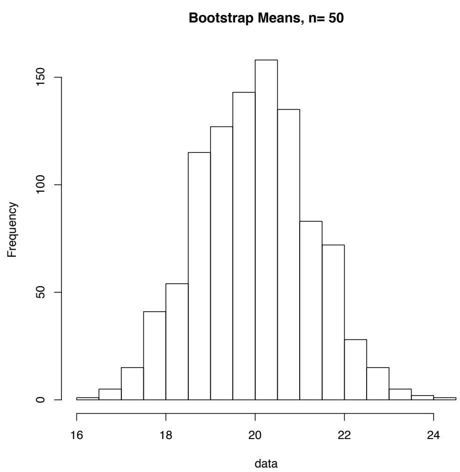
b)

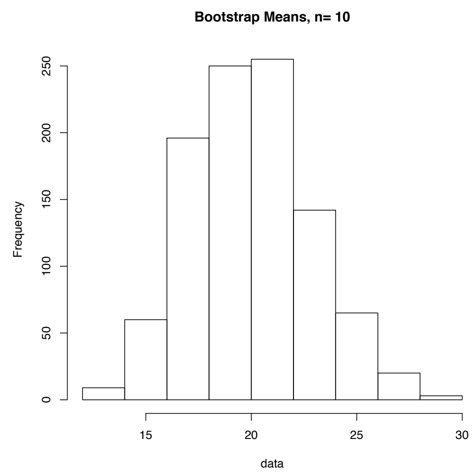
n	Mean	SD
200	20.00	8.94
50	19.98	8.98
10	19.98	9.00

c)

n	Mean	SE
200	20.00	0.65
50	19.98	1.25
10	19.98	2.82



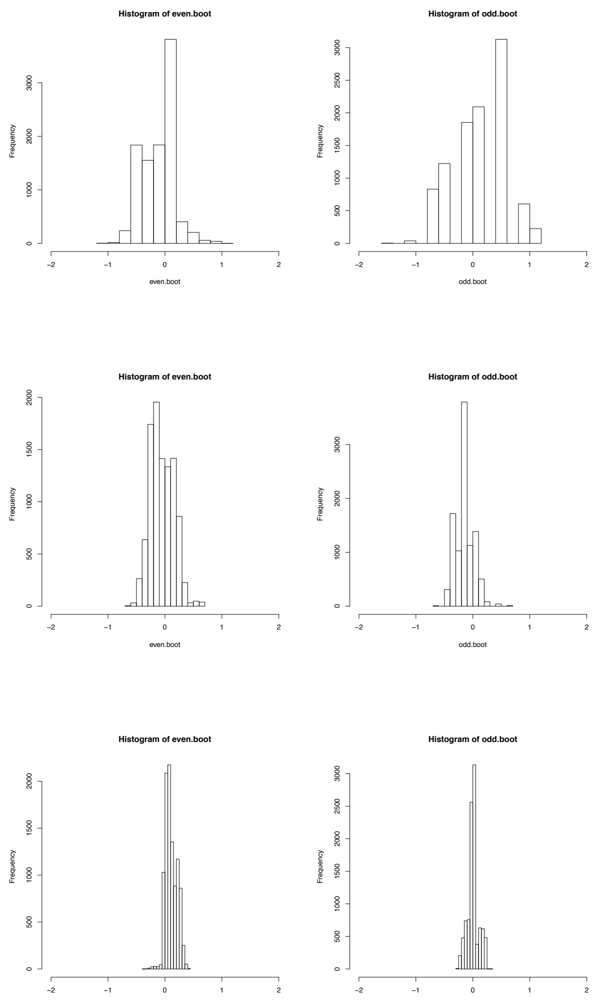




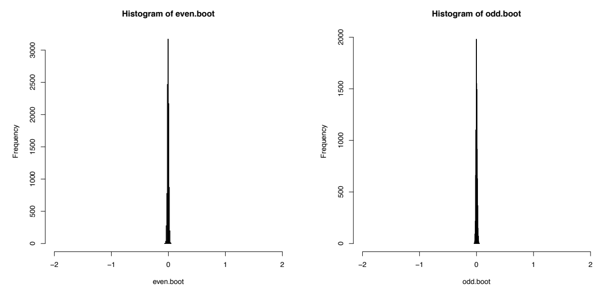
e)

Increasing sample size reduces the bootstrap standard error.

9

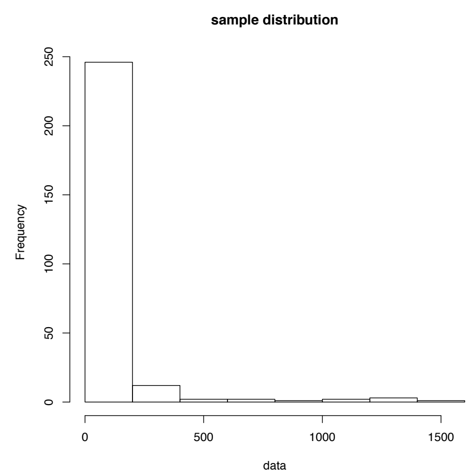


The evenness of N does not seem to affect the histograms, but increasing N reduces the spread.



10

a)

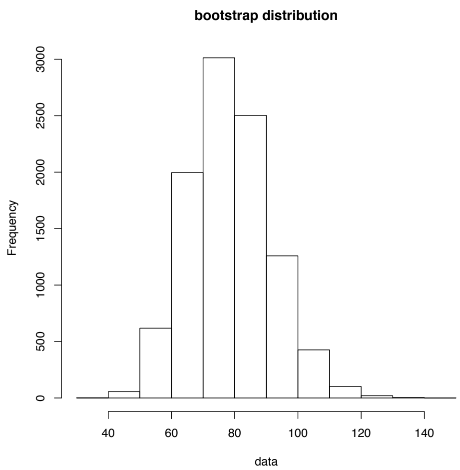


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	5.00	14.20	78.08	55.50	1550.00

The data appears to be highly concentrated between 0 and 500, with several outliers that skew the mean.

b)

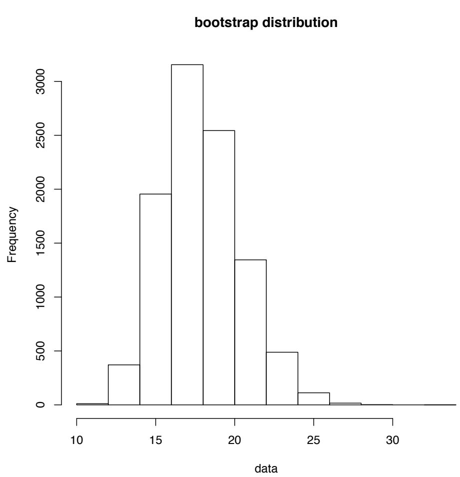
Mean: 78.03
95% confidence interval: 55.0 – 104.61



c)

Bias is 0.37, which is ≈ 3% of the SE.

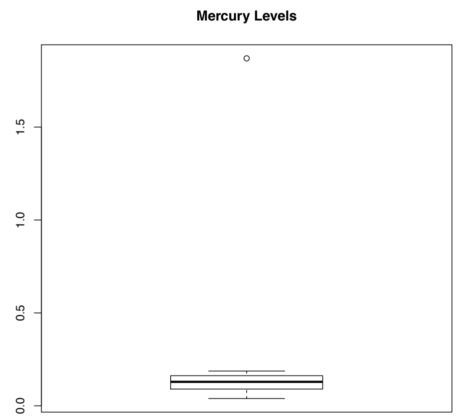
11



Trimmed Mean: 17.90
95% confidence interval: 13.67 – 23.30
Bias is 0.27, which is ≈ 11% of the SE. The new test statistic has a smaller 95% confidence interval at the cost of higher bias.

12

a)



The data is concentrated between 0 and 0.25, with one outlier at 0.9.

b)

Mean: 0.18 ± 0.05
95% Confidence Interval: 0.11–0.31

c)

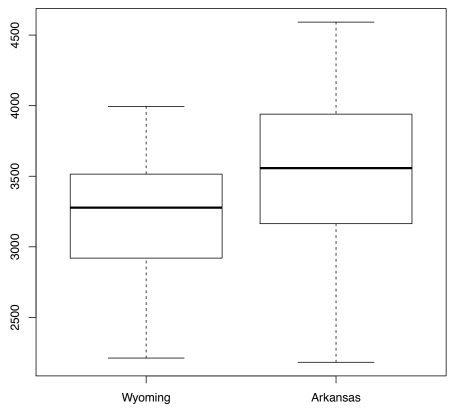
Mean: 0.19 ± 0.06
95% Confidence Interval: 0.11–0.31

d)

Removing the outlier had little effect on the standard error because it was only one point.

14

a)



State	Min.	25%	50%	Mean	75%	Max.
WY	2212	2934	3278	3208	3515	3995
AK	2182	3170	3558	3516	3926	4592

b)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-109.2	232.4	306.3	306.8	381.2	755.9

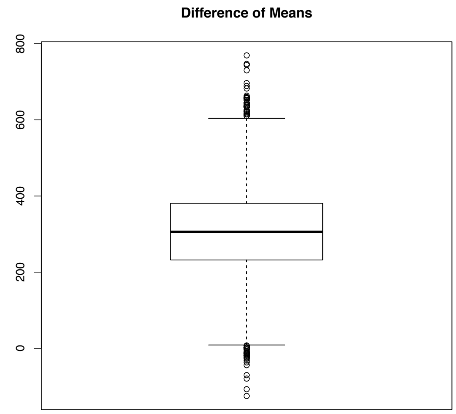
95% confidence interval: 88 – 525

c)

The bias is -0.64, which is about half a percent of the standard error.

d)

Our null hypothesis is that the distribution of birth weights is independent of state. Running the permutation test gave a p-value of 3.5%



e)

We can reject the null at the 5% significance level.