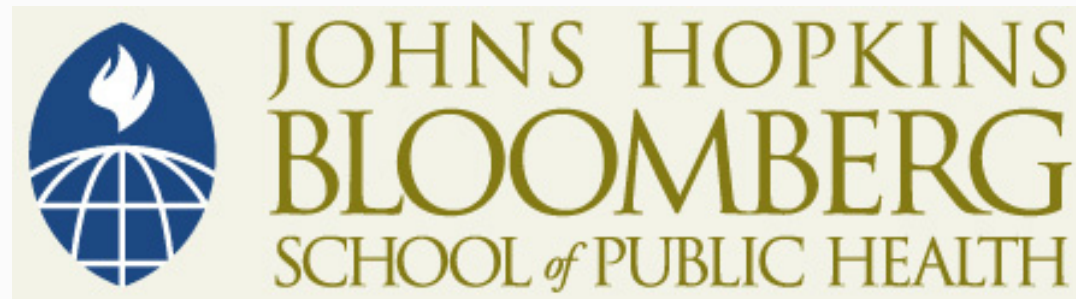


This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2009, The Johns Hopkins University and John McGready. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.



JOHNS HOPKINS  
BLOOMBERG  
SCHOOL *of* PUBLIC HEALTH

## Section C

---

Normal Scores and Variability in Non-normal Data

# Why Do We Like the Normal Distribution So Much?

- The truth is, there is nothing “special” about standard normal scores
  - These can be computed for observations from any sample/ population of continuous data values
  - The score measures how far an observation is from its mean in standard units of statistical distance

# Why Do We Like The Normal Distribution So Much?

- However, unless population/sample has a well known, “well behaved” (like a normal) distribution, we may not be able to use mean and standard deviation to create interpretable intervals, or measure “unusuality” of individual observations

# Hospital Length of Stay Example

- Random sample of 500 patients
  - Mean length of stay: 4.8 days
  - Median length of stay: 3 days
  - Standard deviation: 6.3 days

- Data in Stata

```
list hospstay in 1/10
```

```
+-----+
| hospstay |
+-----+
1. |      2 |
2. |      7 |
3. |      4 |
4. |      5 |
5. |      6 |
+-----+
6. |      5 |
7. |      1 |
8. |      1 |
9. |      1 |
10. |      1 |
+-----+
```

# Hospital Length of Stay Example

- Random sample of 500 patients
  - Mean length of stay: 4.8 days
  - Median length of stay: 3 days
  - Standard deviation: 6.3days

```
. summarize hospstay
```

Variable	Obs	Mean	Std. Dev.	Min	Max
-----+-----					
hospstay	500	4.808	6.282521	1	60

# Hospital Length of Stay Example

- Summarize command with detail option

```
summarize hospstay, detail
```

```
-----  
                    hospstay  
-----  
Percentiles      Smallest  
 1%                1  
 5%                1  
10%               1  
25%               1  
  
50%                3  
                    Largest  
75%                5  
90%               11  
95%               17  
99%               35  
  
Obs                500  
Sum of Wgt.        500  
  
Mean                4.808  
Std. Dev.           6.282521  
  
Variance            39.47008  
Skewness            3.622325  
Kurtosis            21.68121
```

# Hospital Length of Stay Example

- Summarize command with detail option

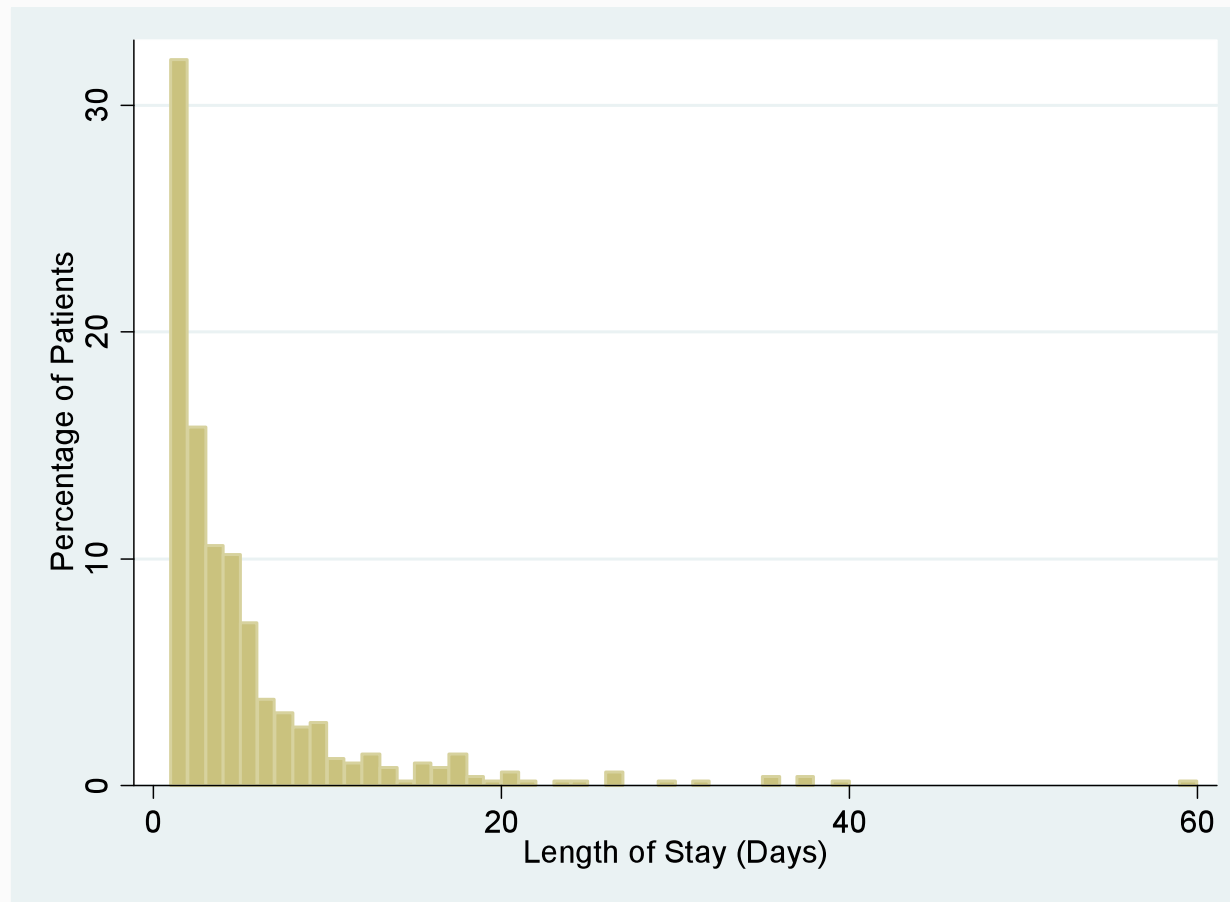
```
summarize hospstay, detail
```

hospstay					
-----					
	Percentiles	Smallest			
1%	1	1			
5%	1	1			
10%	1	1	Obs		500
25%	1	1	Sum of Wgt.		500
50%	3		Mean		4.808
		Largest	Std. Dev.		6.282521
75%	5	37			
90%	11	37	Variance		39.47008
95%	17	39	Skewness		3.622325
99%	35	60	Kurtosis		21.68121



# Hospital Length of Stay Example

- Histogram of sample data

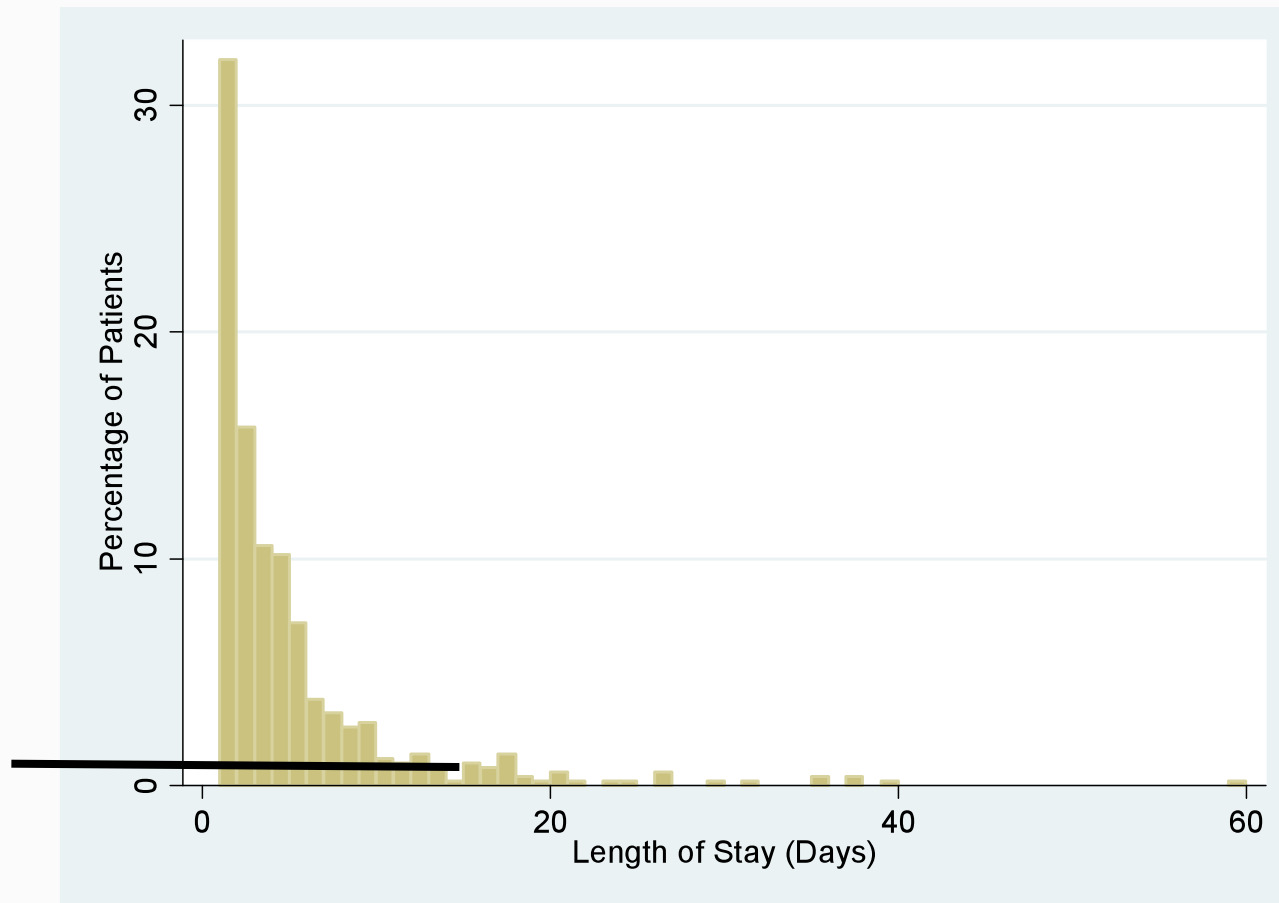


# Constructing Intervals

- Suppose I wanted to estimate an interval containing roughly 95% of the values of hospital length of stay in the population
- Distribution right skewed—can not appeal to properties/methods of normal distribution!
- Mean  $\pm$  2SDs
  - $4.8 \pm 2 \times 6.3$
  - This gives an interval from -7.8 to 17.4 days!

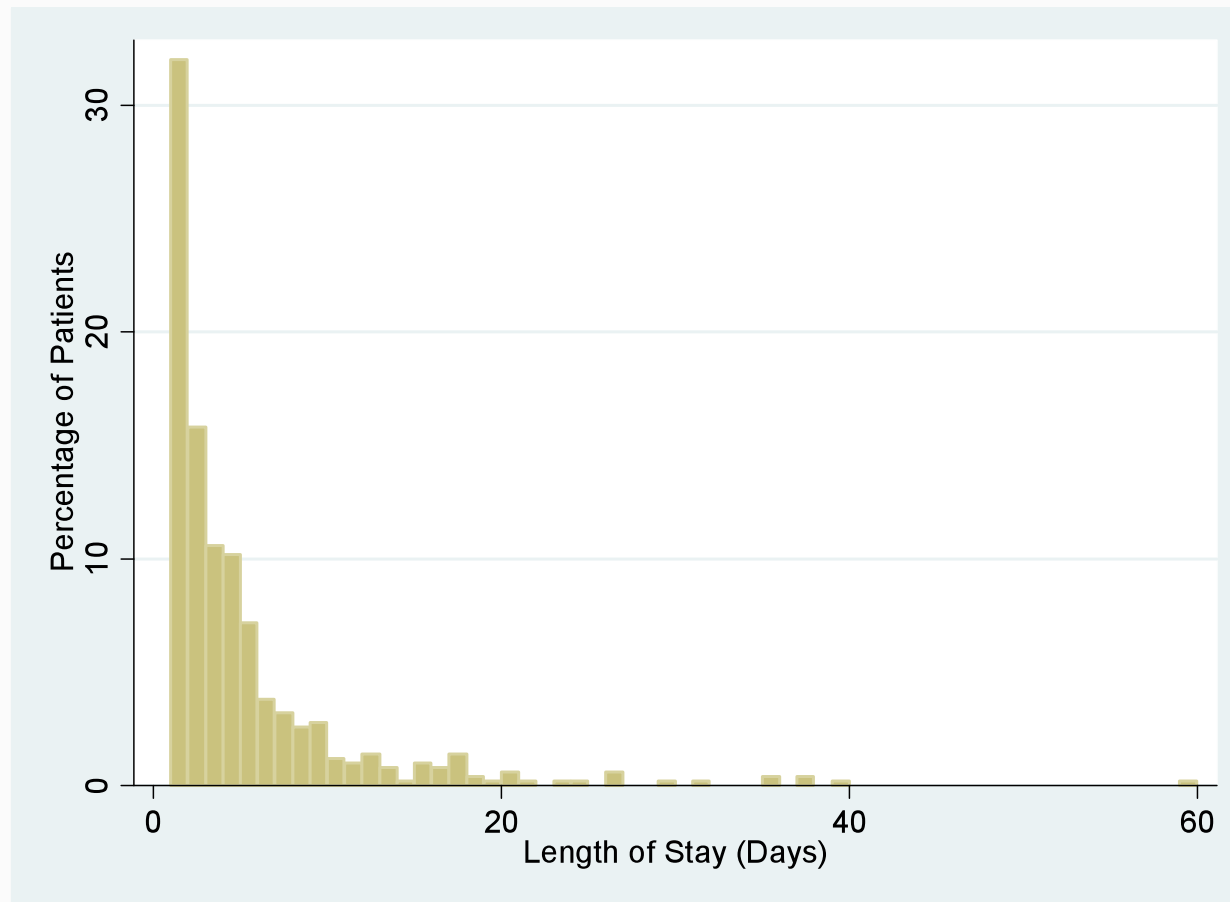
# Hospital Length of Stay Example

- Histogram of sample data



# Constructing Intervals

- We would need to estimate this interval from the histogram and/or by finding sample percentiles



# Constructing Intervals

- Using percentiles
  - Syntax “*centile varname, c(#1, #2, . . .)*”

```
. centile hospstay, c(2.5,97.5)
```

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [ 95% Conf. Interval]	
hospstay	500	2.5	1	1	1
		97.5	23.475	17.69772	32.67554

# Constructing Intervals

- Using percentiles
  - Syntax “*centile varname, c(#1, #2, . . .)*”

```
. centile hospstay, c(2.5,97.5)
```

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [ 95% Conf. Interval]	
hospstay	500	2.5	1	1	1
		97.5	23.475	17.69772	32.67554

- So based on this sample data we estimate that 95% of discharged patients had length of stay between 1 and 24 days

# Constructing Intervals

- What percentage of patients had length of stay greater than five days?

- (Wrong approach) z-score  $z = \frac{5 - 4.8}{6.4} = 0.03$

- Assuming normality, this would suggest that nearly 50% of the patients had length of stay greater than five days

# Hospital Length of Stay Example

- According to percentiles, five days is the 75th percentile: so only 25% of the sample have length of stay over 5 days

summarize hospstay, detail

hospstay						
Percentiles			Smallest			
1%	1	1	1			
5%	1	1	1			
10%	1	1	1	Obs	500	
25%	1	1	1	Sum of Wgt.	500	
50%	3			Mean	4.808	
			Largest	Std. Dev.	6.282521	
75%	5		37	Variance	39.47008	
90%	11		37	Skewness	3.622325	
95%	17		39	Kurtosis	21.68121	
99%	35		60			