Biost 517 / Biost 514 Applied Biostatistics I / Biostatistics I

Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 4:

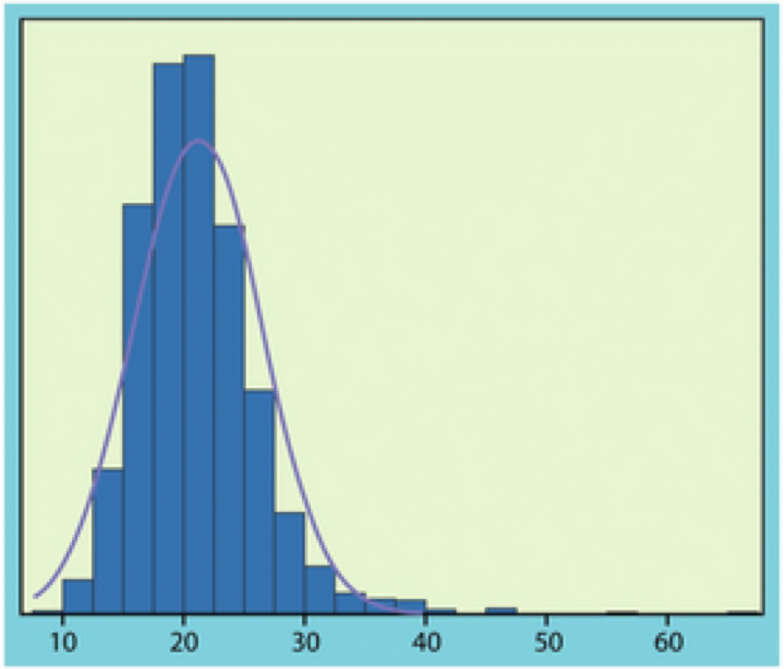
Density Functions; Normal Distribution;

Central Limit Theorem

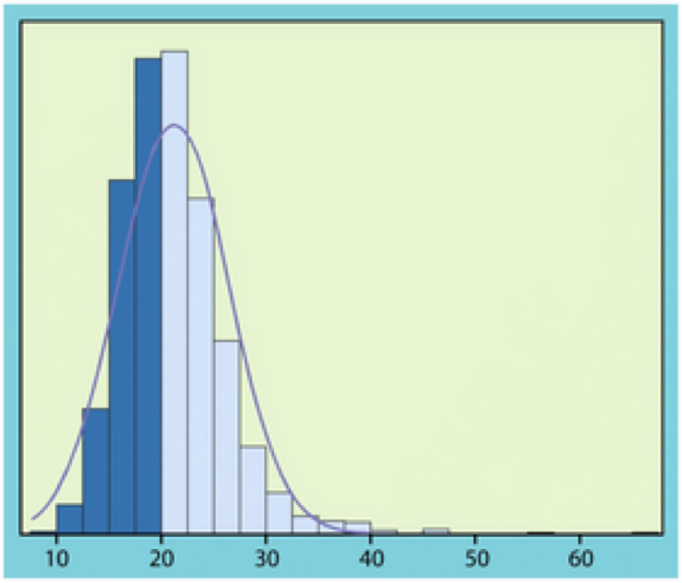
Density Functions

- What's in our toolkit so far?
  - Graphical Display of Quantitative Variable: histograms and boxplots
  - Overall pattern (symmetric, skewed)
  - identify deviations and outliers
  - Descriptive statistics for
    - central tendencies
    - percentiles
    - spread
    - range
- **A new idea**: If the data distribution pattern of a quantitative variable (e.g., histogram) is sufficiently regular, approximate it with a smooth function called a **density function**.

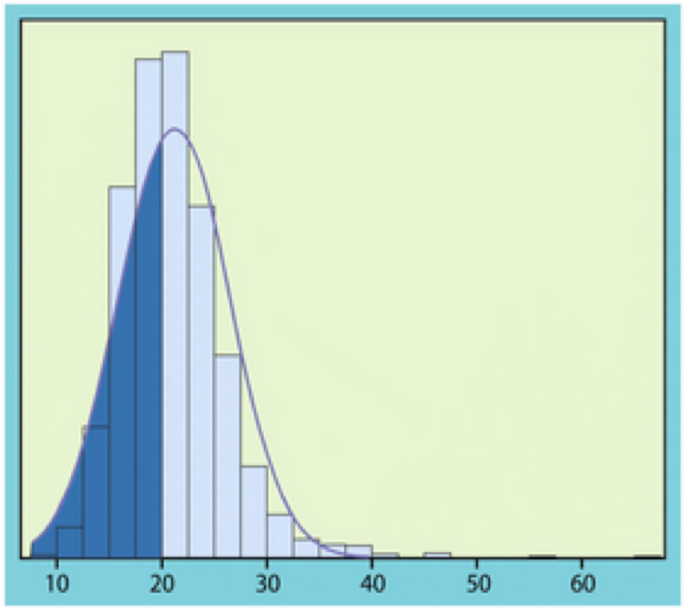
Approximating Data Distributions with Density Functions



True area of data distribution

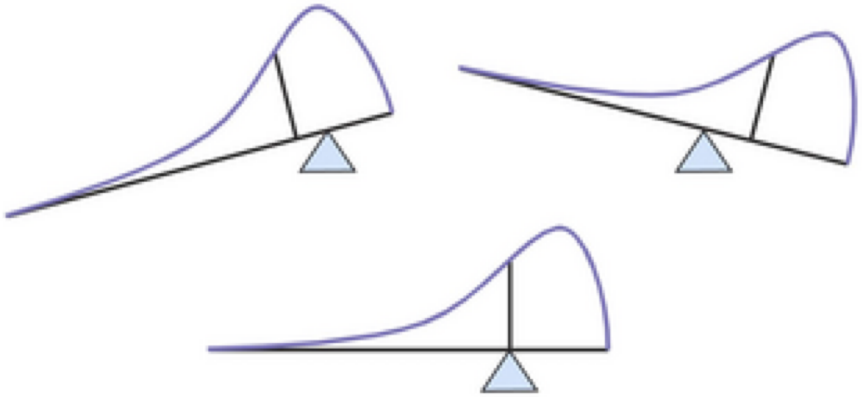


Approximate area using density curve



Mean and Median of Density Curve

- Mean: The balancing point of the density curve of the distribution function, if it were a solid mass.



- Median: The equal-areas point with 50% of the “mass’” on either side of the density curve.
- The mean and median of a symmetric density curve are equal.
- The mean of a skewed density curve is pulled away from the median in the direction of the long tail.

Density Functions

- A **density function** *f(x)* for a continuous random variable *X* has the following properties:
  - *f(x)* ≥ 0, i.e., always on or above the horizontal axis
  - *f(x)* is piecewise continuous
  - total area underneath density function is 1**:**

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

- The density function can be used to obtain frequencies of arbitrary intervals. If *X* is a random variable with density function *f(x)*, then for any *a* < *b,* the probability that X falls in the interval *(a,b) is*

$$P(a \le X \le b) = \int_{a}^{b} f(x)\,dx$$

Cumulative Distribution Functions

- The **cumulative distribution function** (cdf) *F(x)* of a random variable *X* with density function *f(x)* is and is defined as:

$$F(x) = P(X \leq x)$$

- The cdf can be expressed in terms of the density function:

$$F(x) = \int_{-\infty}^{x} f(u)\,du$$

- The cdf can be used to evaluate the probability that *X* falls in an interval:

$$P(a < X < b) = F(b) - F(a)$$

- The cdf *F(x)* is non-decreasing and has values from 0 to 1
- The *p*th quantile of distribution F is the value x_p such that $F(x_p) = p$.
  - Special cases are the quartiles: Q1 with *p=.25, Q2* or median with *p=.5,* and Q3 with *p=.75*

Mean and Variance of Distribution

- For a continuous random variable *X* with density function *f(x)*, the expected value or mean of *X*, *E(X)*, is often denoted as μ and is calculated as follows:

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)\,dx$$

- The variance of *X*, *Var(X)*, is often denoted by σ^2 and is calculated as follows:

$$\sigma^2 = E\left(\left(X - E(X)\right)^2\right) = E\left(\left(X - \mu\right)^2\right) = E(X^2 - 2E(X)\mu + \mu^2)$$

$$= E(X^2) - 2\mu^2 + \mu^2 == E(X^2) - \mu^2$$

where $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx$

- The standard deviation of a *X* is denoted as σ and is the square root of the variance of the distribution:

$$\sigma = \sqrt{\sigma^2}$$

Rules of Expectation and Variances

- Note the following rules for expectation:
  - for two variables *X* and *Y*, E(*X*+*Y*)=E(X)+E(Y)
  - for a constant c, E(c)=c
  - for a constant c and variable *X*, Var(c*X*) = c^2 Var(*X*)
  - If X and Y are two independent variables:
    - Var(*X*+*Y*) = var(*X*)+var(*Y*)
  - In *X* and *Y* are dependent:
    - Var(*X*+*Y*) = Var(*X*)+Var(*Y*)+2Cov(*X*,*Y*) where

$$Cov(X,Y) = E\Big((X - E(X))(Y - E(Y))\Big)$$

Population versus Sample Means and Standard Deviations

- The Greek letters μ and σ are often used to denote the mean and standard deviation, respectively, for a density or a population, to indicate that these are fixed parameters for a population or an idealized model
- In contrast, $\overline{x}$ and s are used to denote the mean and standard deviation, respectively, for a sample.

$$\text{population} \qquad \sigma = \sqrt{E[X - E(X)]^2}$$

$$\text{sample} \qquad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

Densities and Distributions

- It is important to note that density and cumulative distribution functions are only approximations for the distribution of real data, but they simplify analysis and is often accurate enough for practical use.
- There are many widely used continuous distributions
  - Normal
  - Chi Square
  - Uniform
  - Exponential
  - Gamma
  - Student t
  - F

- Densities for distributions come in a variety of shapes with area under the curve in a range of values indicates the proportion of values hat fall in that range.

The Normal Distribution and Density

- The normal distribution is the most widely used distribution is statistics.
- The normal distribution is defined by two parameters: μ (mean), and σ (standard deviation).
- The normal distribution is often written as

$$N(\mu,\sigma^2)$$

- The density function for a normal distribution is

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

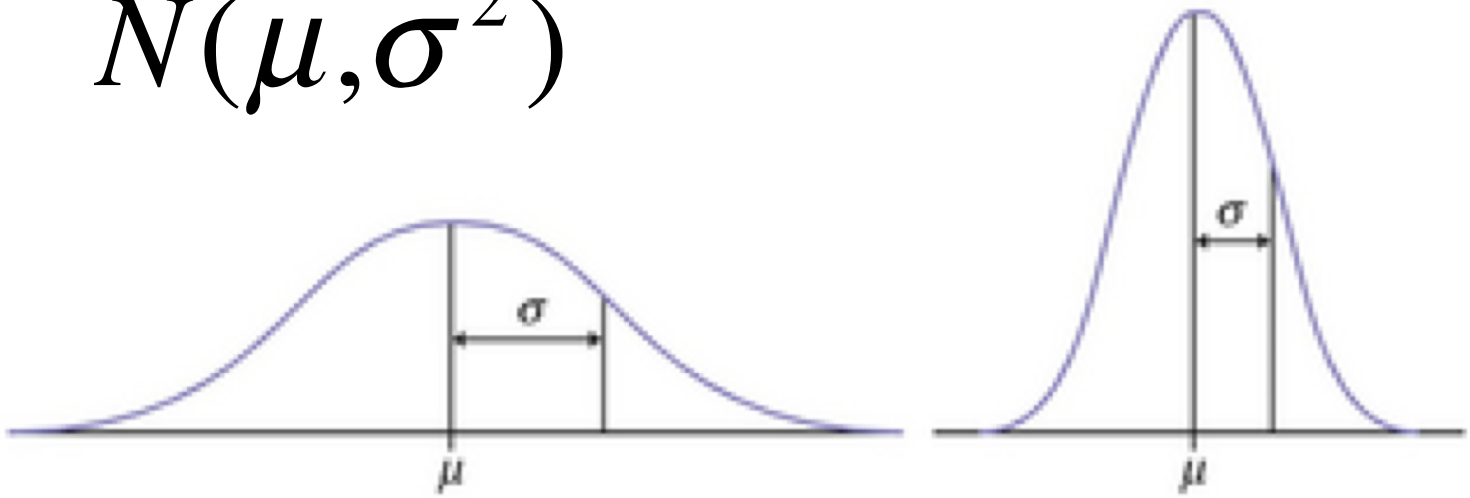
- Can obtain the normal density curve by plotting all possible values for x

The Normal Distribution

- Normal distribution are symmetric, single-peaked, bell-shaped.
- Since the normal distribution is symmetric, μ is also the median

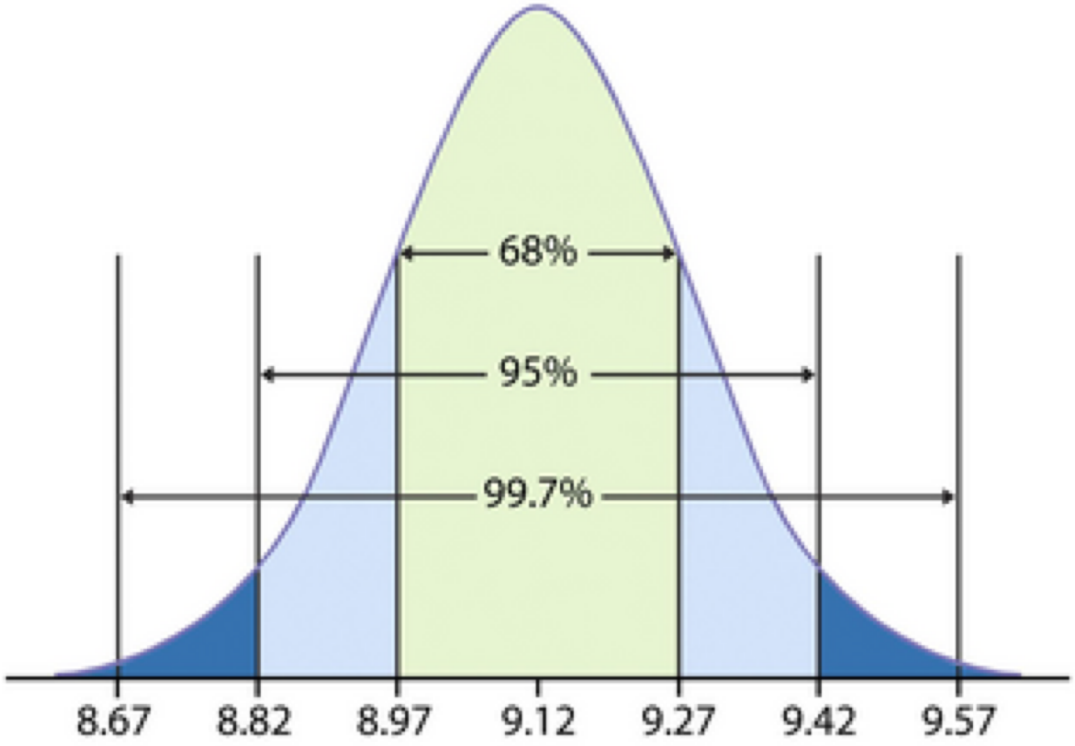
Example density functions

$$N(\mu,\sigma^2)$$



- The point where the curve changes from concave to convex is σ units from μ in either direction.

The 68-95-99.7 Rule: Normal Distributions



- About 68% of the data fall inside $(\mu-\sigma,\mu+\sigma)$
- About 95% of the data fall inside $(\mu-2\sigma,\mu+2\sigma)$
- about 99.7% of the data fall inside $(\mu-3\sigma,\mu+3\sigma)$

Normal Distribution: Example

- Wechsler Adult Intelligence Scale (WAIS) scores are approximately $N\left(110, 25^2\right)$
- About what percent of adults have scores above 110?
- About what percent have scores above 160?
- In what range do the middle 95% of all scores lie?

Standardization and Z-Scores

- All normal distributions are the same if we measure in units of size σ about the mean μ
- If a data point x is from $N(\mu,\sigma^2)$, then a standardized value, or **z-score**, can be obtained by:

$$z = \frac{x - \mu}{\sigma}$$

- A z-score tells us how many standard deviations x is from μ and in what direction.
- The standardized value, or z-score, has density $N(0,1)$, which is referred to as **standard normal.** This enables us to use a standard normal distribution to find probabilities for any normal variable

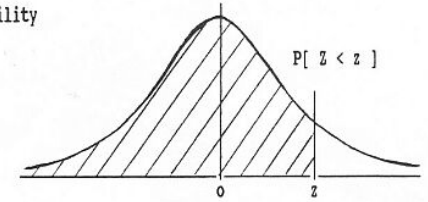
Standard Normal Table

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z

i.e. $$P[\ Z < z\ ] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\, dZ$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

Standard Normal Table

Standard normal calculations

1. State the problem in terms of x.
2. Standardize: $z = \frac{x-\mu}{\sigma}$.
3. Look up the required value(s) on the standard normal table.
4. Reality check: Does the answer make sense?

Standard Normal Table

Backward Normal Calculations

We can also calculate the values, given the probabilities:

"Backward" normal calculations

1. State the problem in terms of the probability of being **less** than some number.
2. Look up the required value(s) on the standard normal table.
3. "Unstandardize," i.e. solve $z = \frac{x-\mu}{\sigma}$ for x.

$$X = \sigma Z + \mu$$

R: Cumulative Frequencies

- Standard normal tables were necessary until statistical software packages became widely available
- To obtain the cdf for a normal distribution, the ***pnorm()*** function can be used. Use the help function in R or more details: ***?pnorm***

- For example, to obtain the probability that a random variable X has a value less than 2, where X is from a normal distribution with mean 1 and standard deviation 1, the following command can be used:

```
> pnorm(2,mean=1,sd=1)
[1] 0.8413447
```

- How would you obtain this with the standard normal table?
  - The standardized z-score is (2-1)/1=1
  - From the standard normal table, the area to the left of 1 is 0.8413

R: Normal Quantiles

- To obtain quantiles/percentiles for a normal distribution in R, you only need to use the **qnorm()** function. Use the help function in R or more details: ***?qnorm***
- For examples, to obtain the 95th percentile for a normal distribution with mean 140 and standard deviation 30, the following command can be used:

```
> qnorm(.95,mean=140,sd=30)
[1] 189.3456
```

- How would you obtain this answer using the standard normal table?
  - The 95th percentile for a standard variable is around 1.645 from the table
  - Back solving, we have that that the 95th percentile must be

$$x = \sigma z + \mu = (30)(1.645) + 140 = 189.35$$

Mean and Variance of Sample Mean

- Recall that the **sample mean** is defined as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_{n-1} + x_n}{n}$$

where $x_1, x_2, \ldots x_n$ are realized values of some variables $X_1, X_2, \ldots X_n$ in a sample of size *n*

- Let's be more specific about what we mean by a sample of size *n*
- Consider the sample to be a collection of *n* **independent and identically distributed (or iid)** random variables $X_1, X_2, \ldots X_n$ with common mean μ and common standard deviation σ, i.e.,

$$X_i \sim N(\mu, \sigma^2) \text{ for each } i,$$

- Let's derive the mean and variance of

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Expected Value of Sample Mean

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n}\left(n\right)\mu = \mu$$

- Across a large number of replicated studies, the average of all of the sample means from the studies will be close to the true mean in the population

Variance and SE of Sample Mean

- The variance of $\bar{X}$:

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i\right) = \frac{1}{n^2}\left(n\right)\sigma^2 = \frac{\sigma^2}{n}$$

- The standard deviation of $\bar{X}$, also called its **standard error (SE)**, is

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Variability of the Sample Mean

- The properties of variance help us describe the variability of statistics, such as the sample mean, computed across many replicate experiments
- Suppose we sample *n* independent measurements from a population in which
  - the average value in the population is μ ,
  - the variance in the population is σ^2
- Then, across a large number of replicated studies with a sample size *n*
  - average of the sample means will be around μ
  - variance of the sample means will be around $\frac{\sigma^2}{n}$
  - standard deviation or (SE) of the sample means will be around $\frac{\sigma}{\sqrt{n}}$

Asymptotic Theory for Sample Mean

- Note that in the previous example, we did not assume to know the distribution of the random variables
- Sometimes the exact sampling distribution of a statistic is unknown or too difficult to compute
- We usually don't know (and often don't want to assume) the distribution of the variable in the population
- Theoretical results can allow for statistical inference about a parameter when sample sizes are sufficiently large, and most of this **asymptotic theory** is based on some form of a **Central Limit Theorem** for the distribution of a sum or arithmetic mean of random variables
  - **Asymptotics** is when the sample size becomes infinite
- The CLT tells us about the sampling distribution of the parameter estimate, even when the true distribution of the data is unknown

Central Limit Theorem (CLT)

The Central Limit Theorem

Now we know that $\bar{X}$ has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, but what is its distribution?

If $X_1, X_2, \ldots, X_n$ are *Normally distributed*, then $\bar{X}$ is also normally distributed. Thus,

$$X_i \sim N(\mu, \sigma^2),\ \forall i \implies \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

If $X_1, X_2, \ldots, X_n$ are *not* Normally distributed, then the Central Limit Theorem tells us that $\bar{X}$ is *approximately* Normal.

The Central Limit Theorem

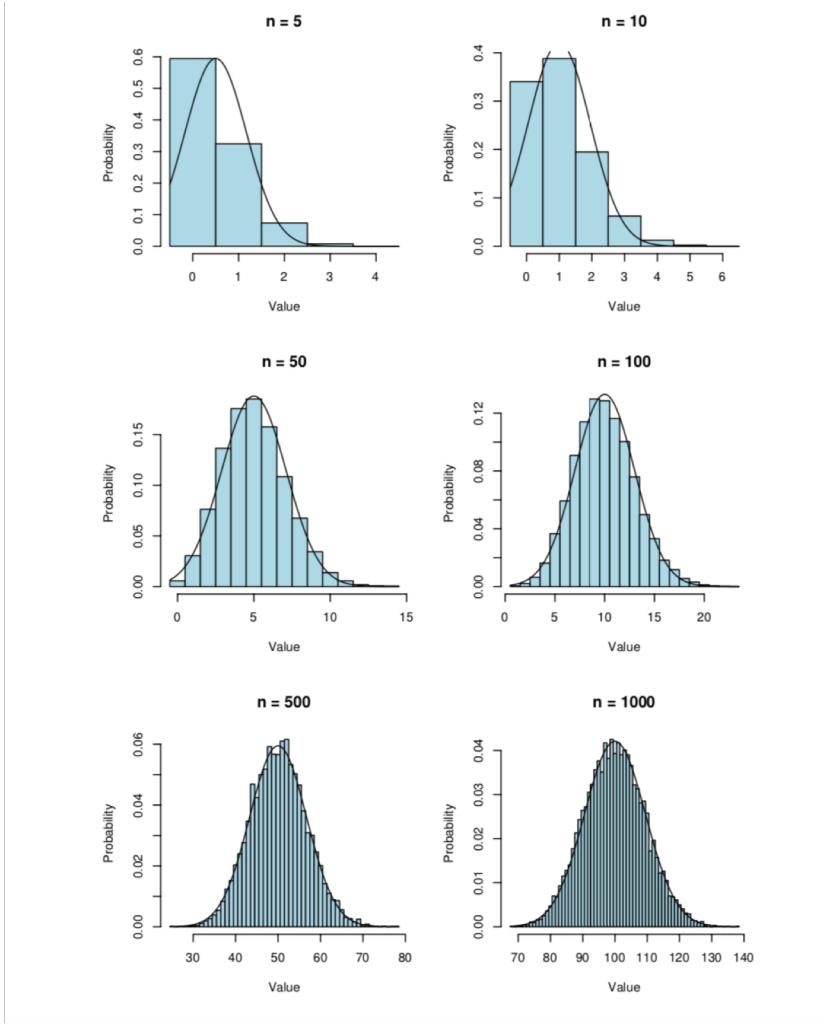
Suppose $X_1, X_2, \ldots, X_n$ are *iid* random variables with mean μ and finite standard deviation σ.
If n is sufficiently large, the sampling distribution of $\bar{X}$ is approximately Normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

CLT and Sample Size

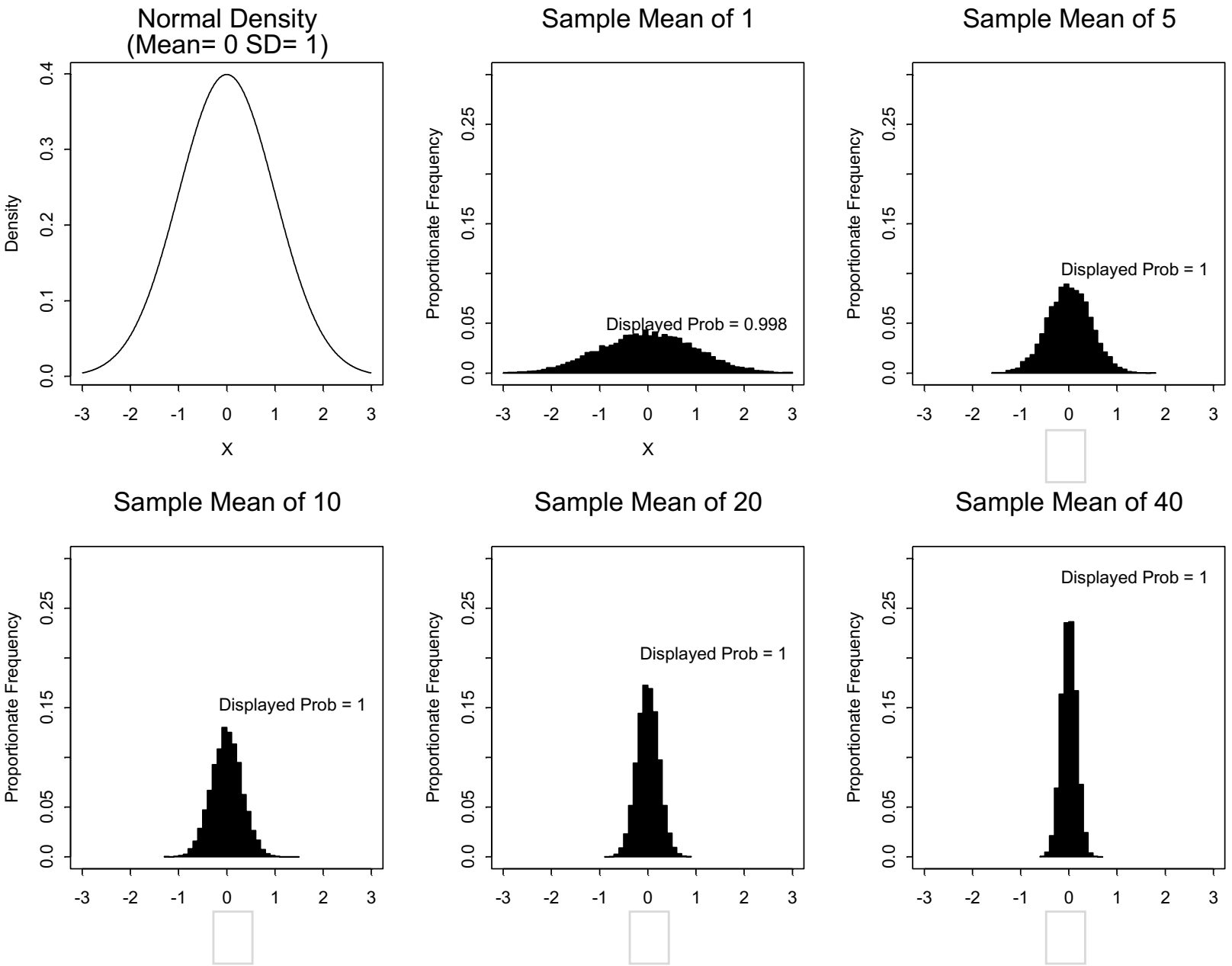
- The definition of a sufficiently "large" sample depends on the original distribution
- Distributions with heavy tails or extreme skewness, require larger sample sizes for a good approximation for normality of the sample mean
- The definition of "large" sample size depends on how extreme the probabilities that you want
  - Accuracy for the 25th or 75th percentile is better than for the 1st or 99th percentile
- That said, it is often surprising how small "large" is
  - A sample size of 30 – 40 will suffice for the Central Limit Theorem most non-normal data sets commonly encountered in practice
  - (see Lumley, et al., *Annual Rev. Pub Health,* 2002)

CLT Example: Binary Variable

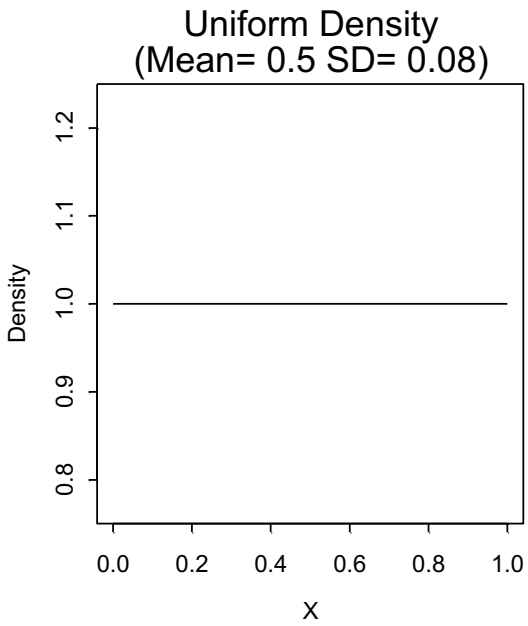
- Histograms, each representing 10,000 samples, from the sum of Bernoulli (binary) random variables with p = 0.1 for different sample sizes

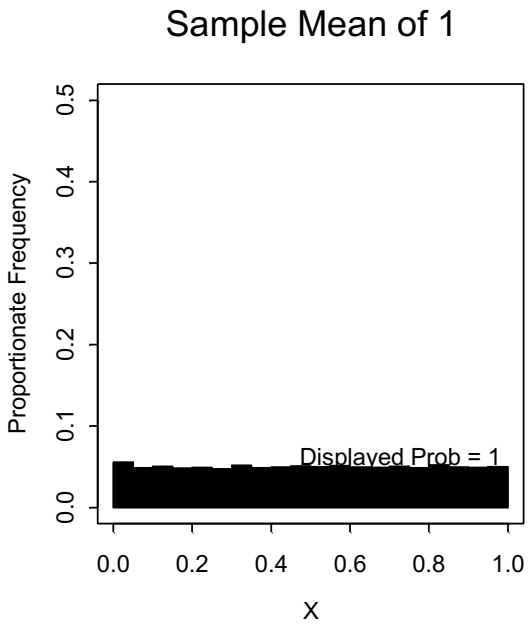


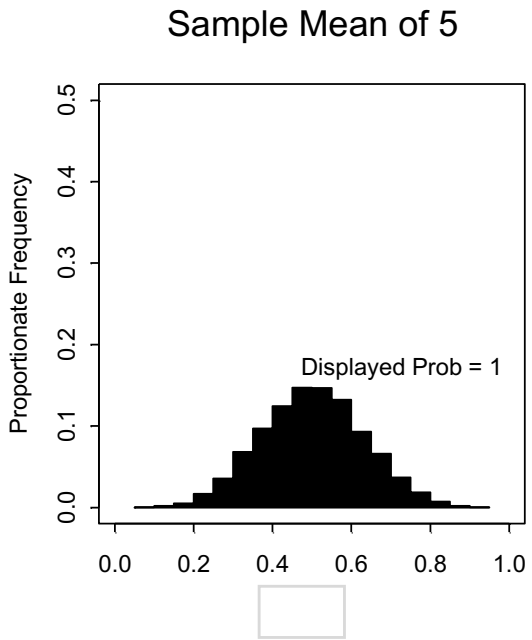
Normal Data

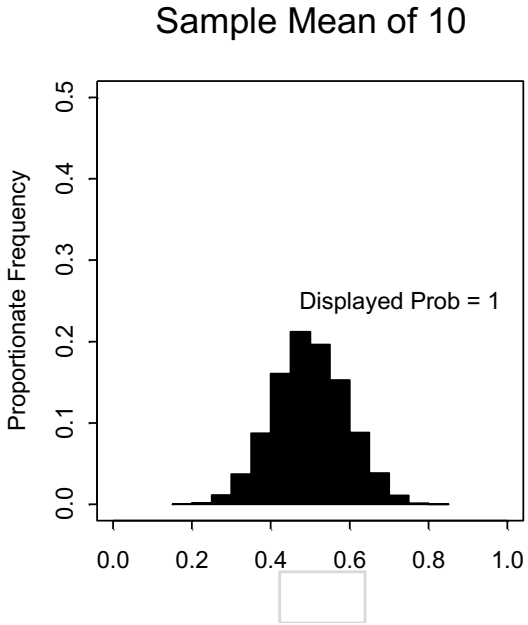

Normal Density
(Mean= 0 SD= 1)
Density
X
Sample Mean of 1
Proportionate Frequency
Displayed Prob = 0.998
X
Sample Mean of 5
Proportionate Frequency
Displayed Prob = 1
Sample Mean of 10
Proportionate Frequency
Displayed Prob = 1
Sample Mean of 20
Proportionate Frequency
Displayed Prob = 1
Sample Mean of 40
Proportionate Frequency
Displayed Prob = 1

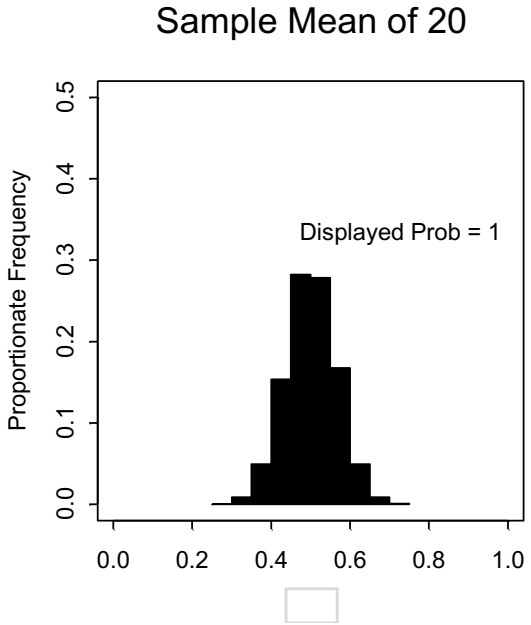
Uniform Data

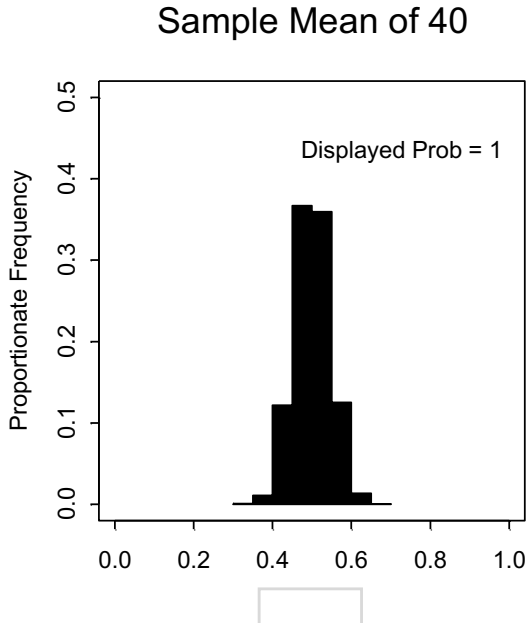




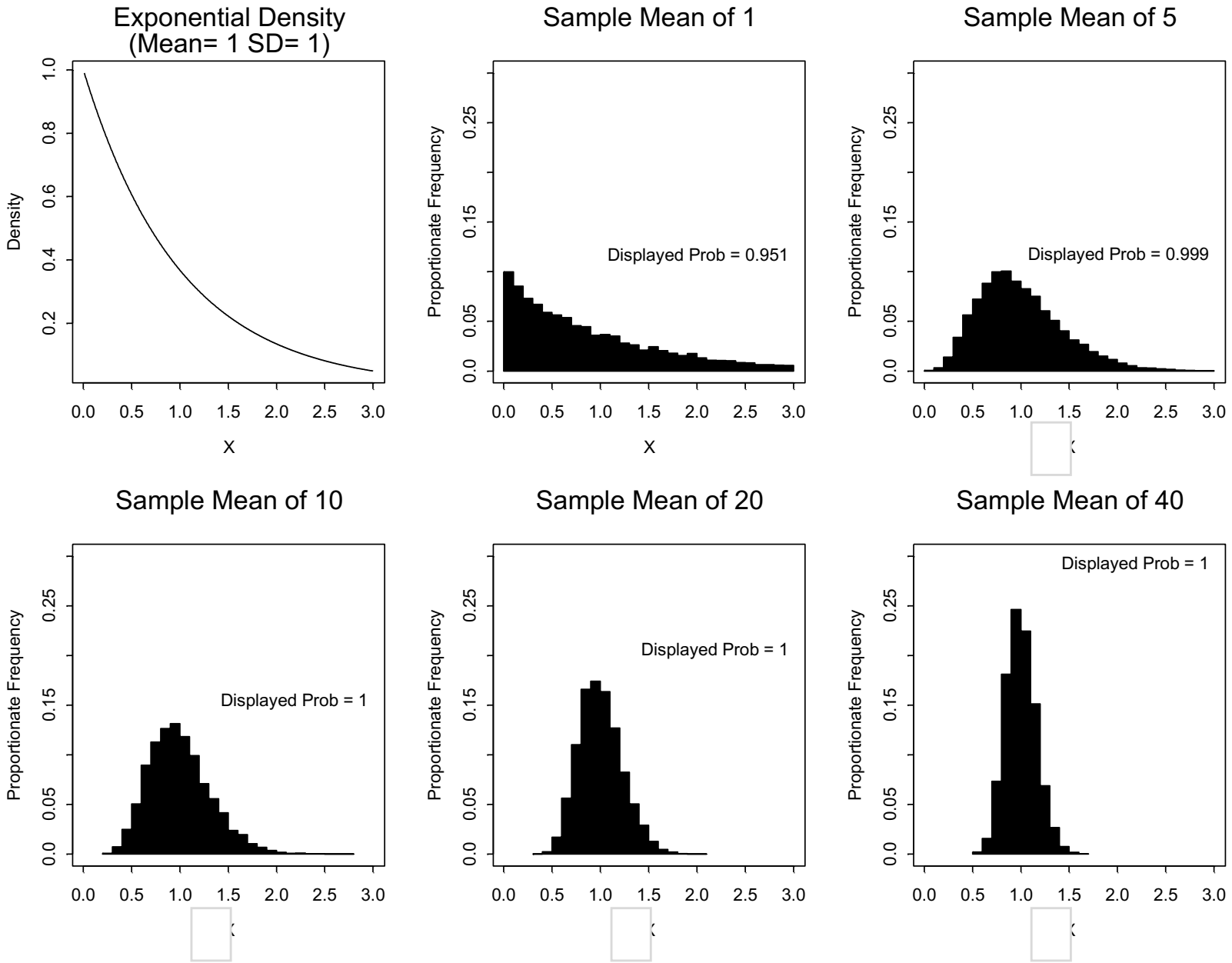




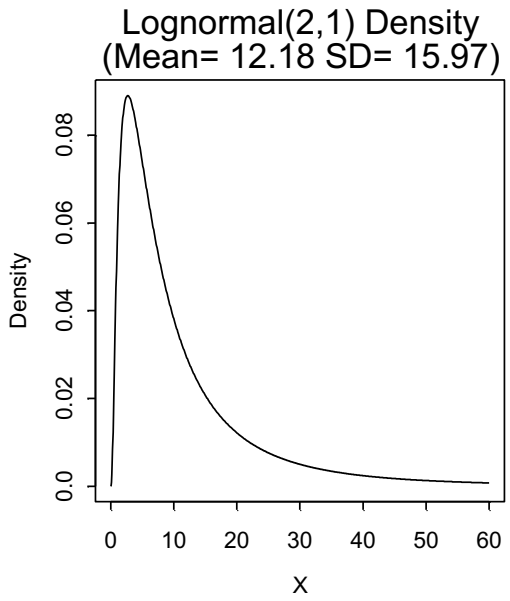


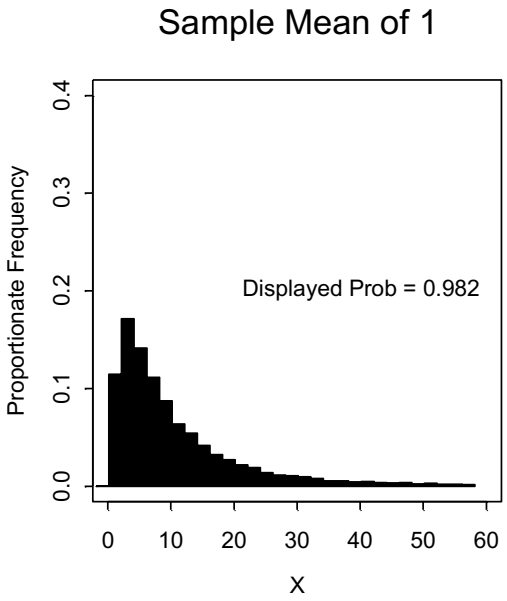


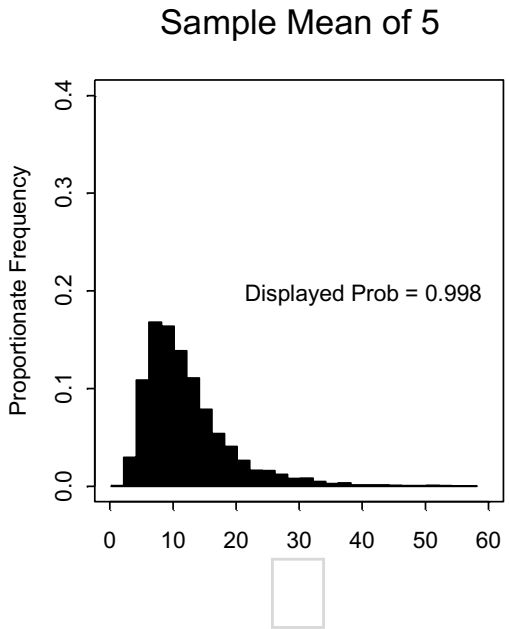
Exponential Data

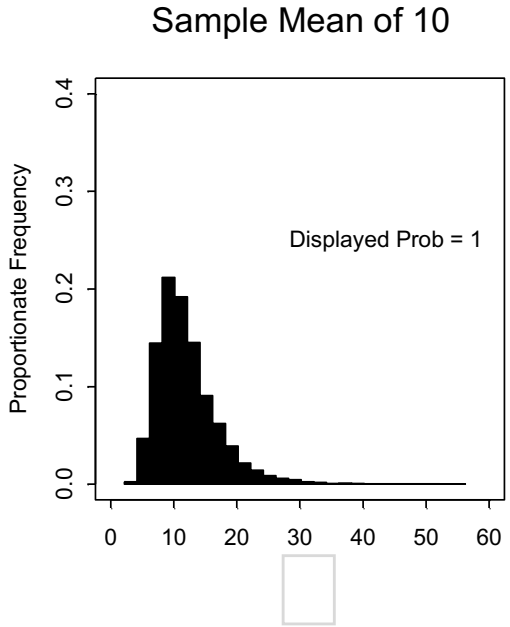


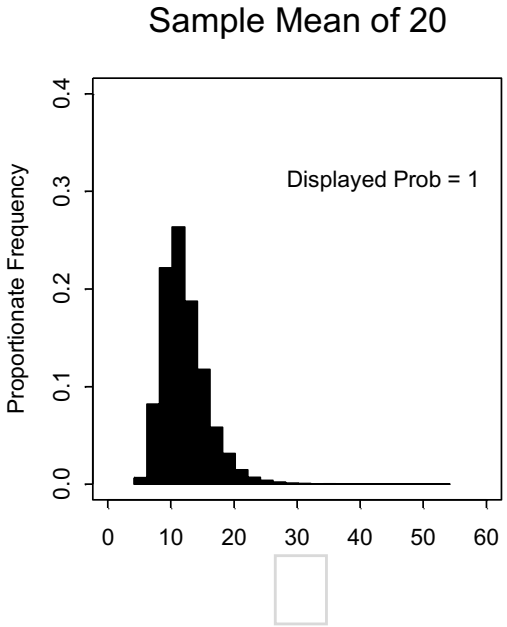
Lognormal Data (Mod. Skewed)

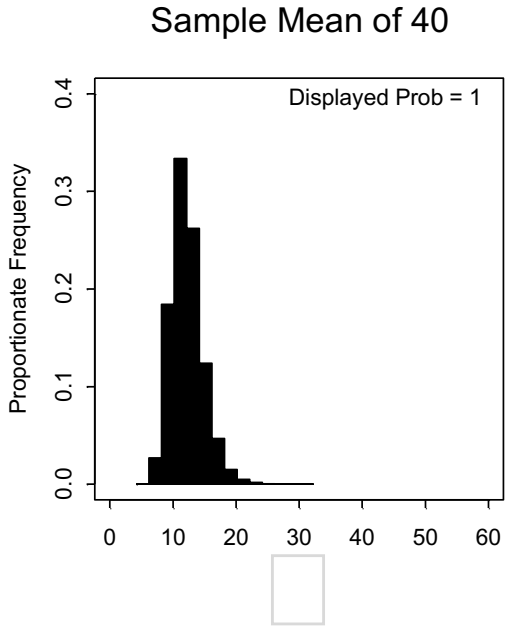




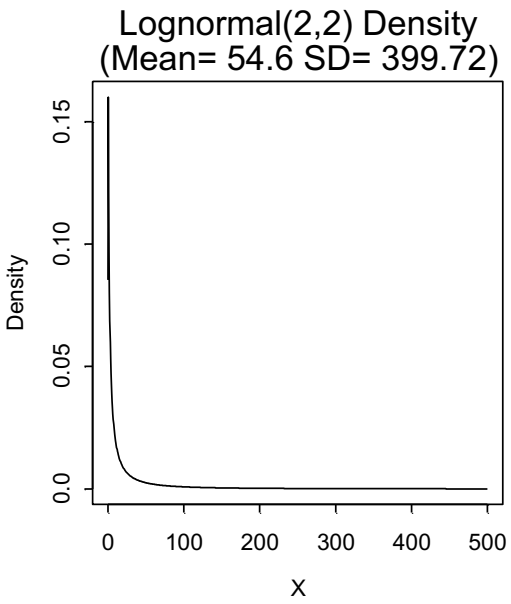


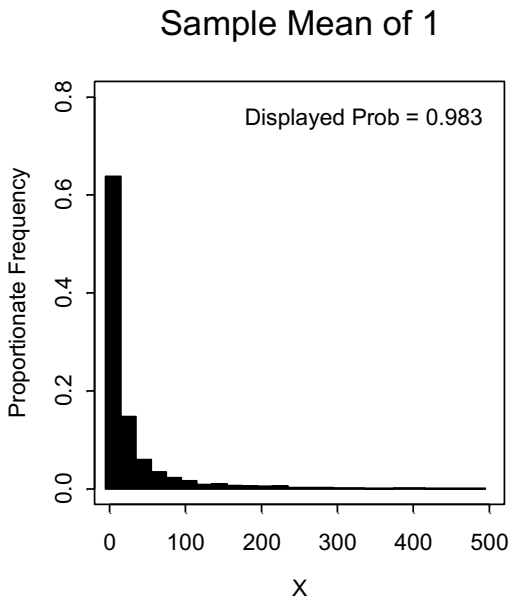


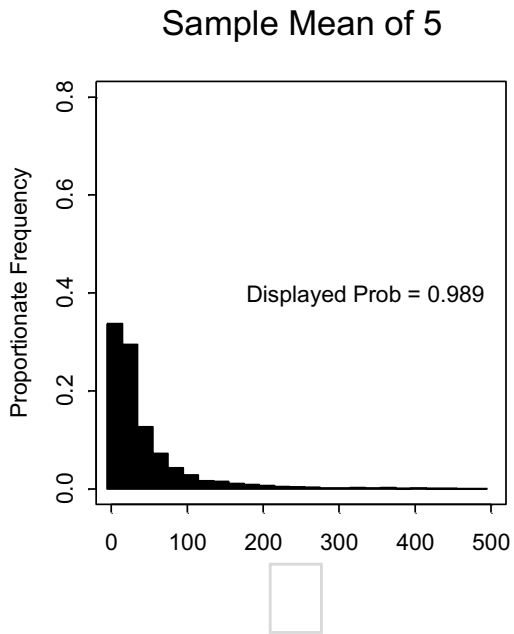


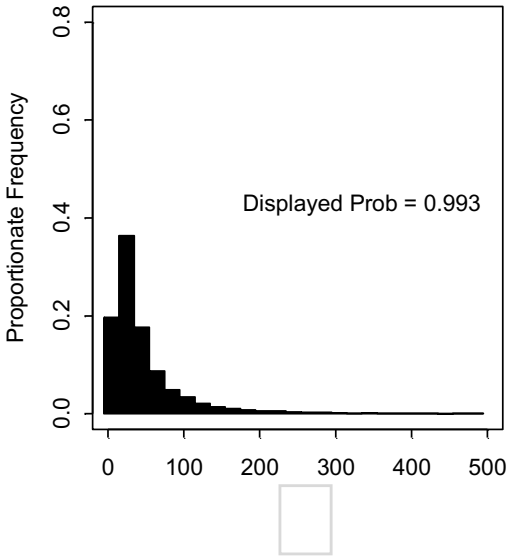


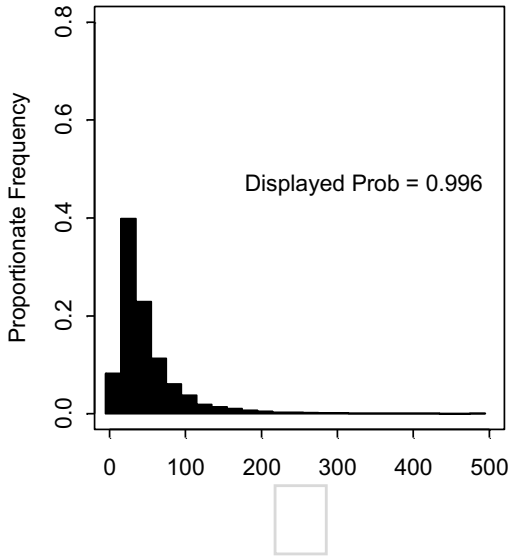
Lognormal Data (Highly Skewed)

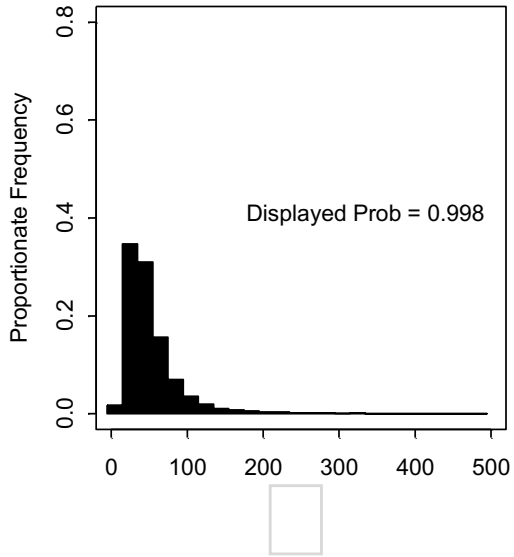




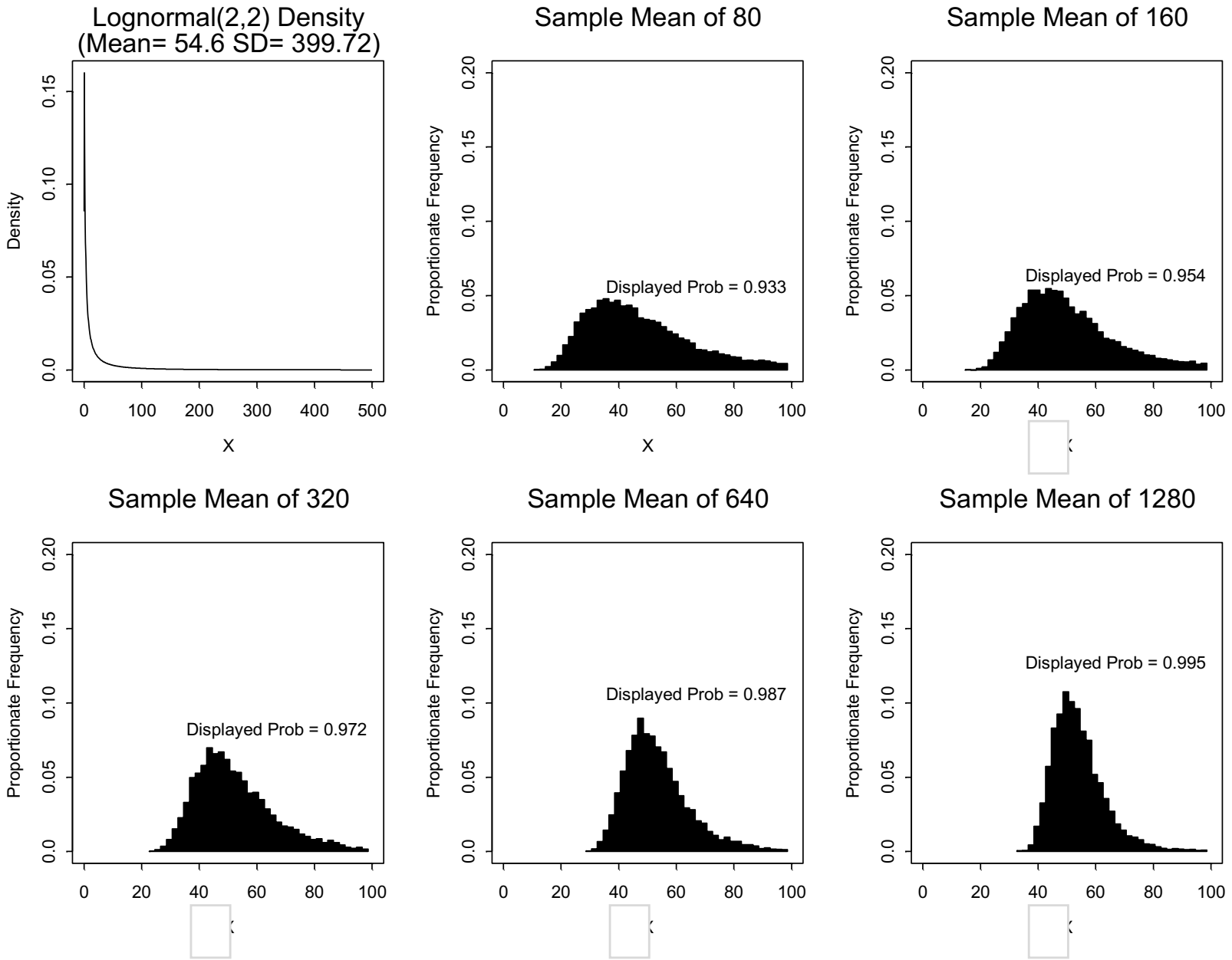








Lognormal Data (Highly Skewed)



Other Central Limit Theorems

- We have the CLT for the mean in the “classical” setting:
  - Independent, identically distributed data from a distribution with finite variance
- “Specialized” CLTs exist for some other settings
  - Independent, but not identically distributed
  - Identically distributed, but correlated data
  - Transformations of sample means
- We will generally leave it to those who write our software to get the formulas right
  - Our job is to know what the CLT says and does not say
  - Our job is to understand what assumptions we make when we analyze data and judge whether they are reasonable