Mathematical statistics

September 11st, 2018

Lecture 5: Sampling distributions

Topics

Week 1	Probability reviews
Week 2	**Chapter 6: Statistics and Sampling Distributions**
Week 4	Chapter 7: Point Estimation
Week 7	Chapter 8: Confidence Intervals
Week 10	Chapter 9: Test of Hypothesis
Week 14	Regression

Overview

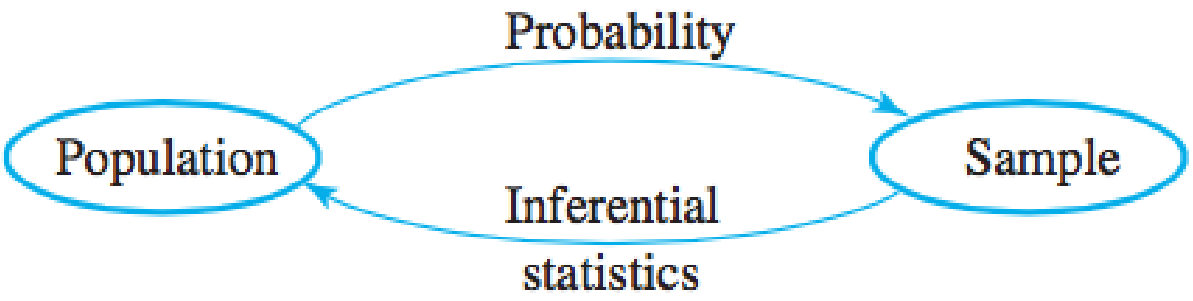
6.1 Statistics and their distributions

6.2 The distribution of the sample mean

6.3 The distribution of a linear combination

Order $6.1 \rightarrow 6.3 \rightarrow 6.2$

Random sample



Definition

The random variables $X_1, X_2, ..., X_n$ are said to form a (simple) random sample of size n if

1. the X_i's are independent random variables
2. every X_i has the same probability distribution

Statistics

Definition

A statistic is any quantity whose value can be calculated from sample data

- prior to obtaining data, there is uncertainty as to what value of any particular statistic will result $\rightarrow$ a statistic is a random variable
- the probability distribution of a statistic is referred to as its *sampling distribution*

Example of a statistic

- Let $X_1, X_2, \ldots, X_n$ be a random sample of size n
- The sample mean of $X_1, X_2, \ldots, X_n$, defined by

$$\bar{X} = \frac{X_1 + X_2 + \ldots X_n}{n},$$

  is a statistic
- When the values of $x_1, x_2, \ldots, x_n$ are collected,

$$\bar{x} = \frac{x_1 + x_2 + \ldots x_n}{n},$$

  is a realization of the statistic $\bar{X}$

Questions for this chapter

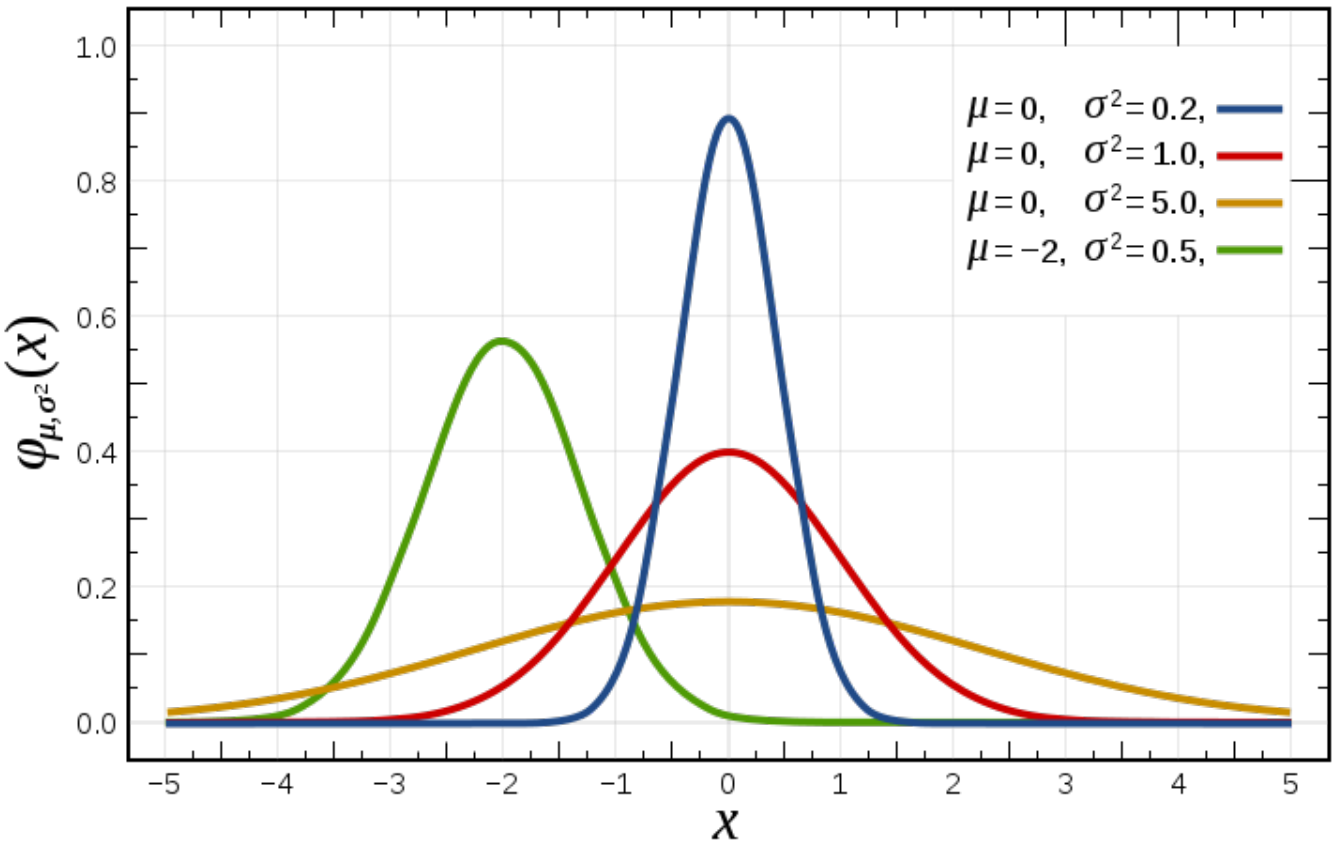
Given a random sample $X_1, X_2, \ldots, X_n$, and

$$T = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$

- If we **know** the distribution of X_i's, can we obtain the distribution of T?
  - Simple cases
  - If $X_i's$ follow normal distribution, then so does T.
- If we **don't know** the distribution of X_i's, can we still obtain/approximate the distribution of T?
  - Can we at least compute the mean and the variance?
  - When T is the sample mean, i.e. $a_1 = a_2 = \ldots = \frac{1}{n}$

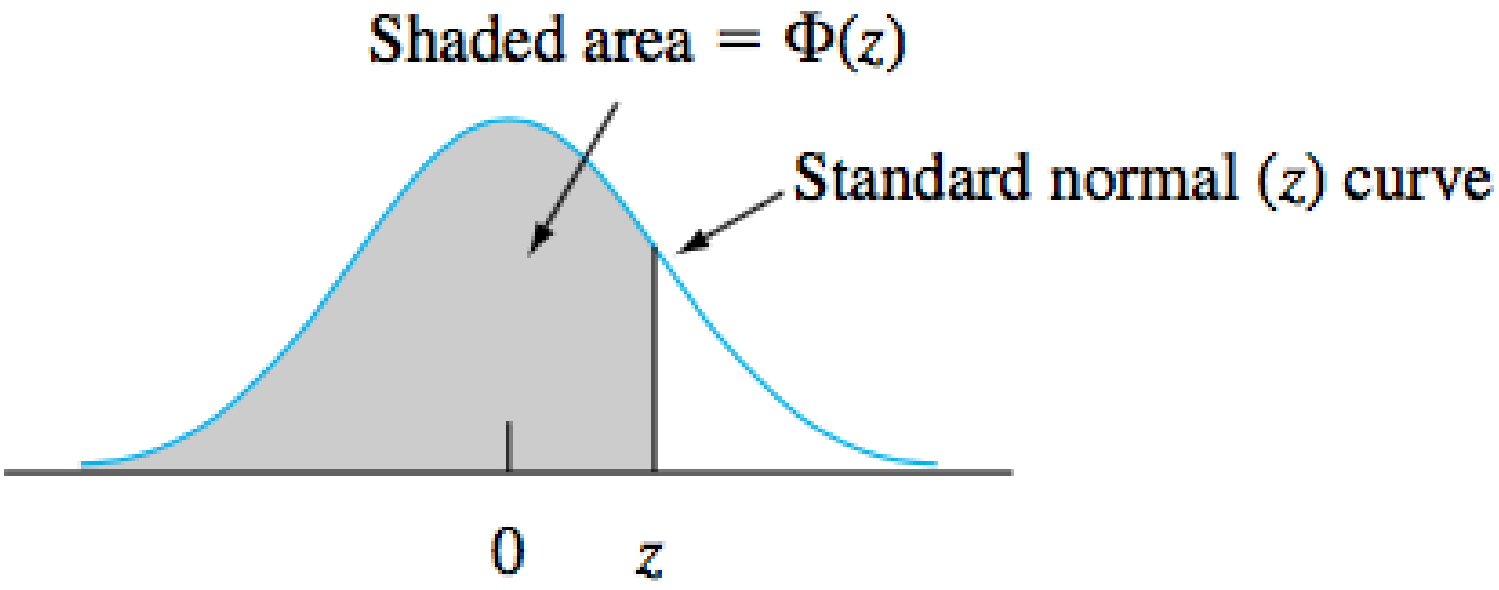
Linear combination of normal random variables

$\mathcal{N}(\mu, \sigma^2)$



$$E(X) = \mu, Var(X) = \sigma^2$$

$\Phi(z)$



$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} f(y)\ dy$$

Table A.3 Standard Normal Curve Areas (*cont.*) $\Phi(z) = P(Z \leq z)$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Shifting and scaling normal random variables

Problem

Let X be a normal random variable with mean μ and standard deviation σ.
Then

$$Z = \frac{X - \mu}{\sigma}$$

follows the standard normal distribution.

Linear combination of normal random variables

Theorem

Let $X_1, X_2, \ldots, X_n$ be independent normal random variables (with possibly different means and/or variances). Then

$$T = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$

also follows the normal distribution.

What are the mean and the standard deviation of T?

- $E(T) = a_1 E(X_1) + a_2 E(X_2) + \ldots + a_n E(X_n)$
- $\sigma_T^2 = a_1^2 \sigma_{X_1}^2 + a_2^2 \sigma_{X_2}^2 + \ldots + a_n^2 \sigma_{X_n}^2$

Example

Problem

Let $X_1, X_2, \ldots, X_{16}$ *be a random sample from* $\mathcal{N}(1, 4)$ *(that is, normal distribution with mean* $\mu = 1$ *and standard deviation* $\sigma = 2$*).*
Let $\bar{X}$ *be the sample mean*

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_{16}}{16}$$

- *What is the distribution of* $\bar{X}$*?*
- *Compute* $P[\bar{X} \leq 1.82]$

Example

Problem

Two airplanes are flying in the same direction in adjacent parallel corridors. At time $t = 0$, the first airplane is 10 km ahead of the second one.
Suppose the speed of the first plane (km/h) is normally distributed with mean 520 and standard deviation 10 and the second planes speed, independent of the first, is also normally distributed with mean and standard deviation 500 and 10, respectively.

What is the probability that after 2h of flying, the second plane has not caught up to the first plane?

What if X_i's are not normal distributions?

Given a random sample $X_1, X_2, \ldots, X_n$, and

$$T = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$

If we don't know the distribution of X_i's, can we obtain the distribution of T?

Linear combination of random variables

Theorem

Let $X_1, X_2, \ldots, X_n$ be independent random variables (with possibly different means and/or variances). Define

$$T = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n,$$

then the mean and the standard deviation of T can be computed by

- $E(T) = a_1 E(X_1) + a_2 E(X_2) + \ldots + a_n E(X_n)$
- $\sigma_T^2 = a_1^2 \sigma_{X_1}^2 + a_2^2 \sigma_{X_2}^2 + \ldots + a_n^2 \sigma_{X_n}^2$

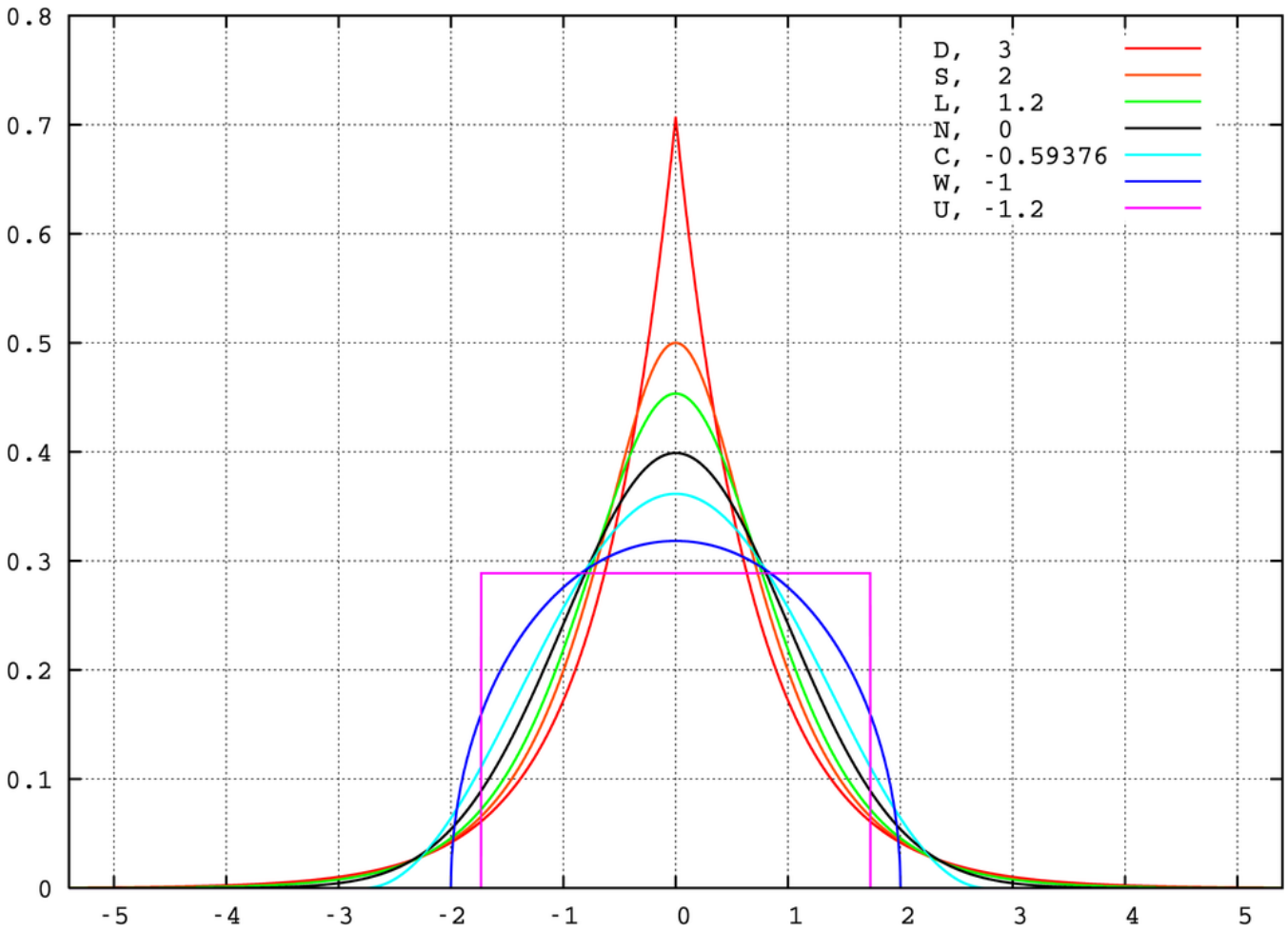
Example

Problem

A gas station sells three grades of gasoline: regular unleaded, extra unleaded, and super unleaded. These are priced at 2.20, 2.35, and 2.50 per gallon, respectively.
Let X_1, X_2, and X_3 denote the amounts of these grades purchased (gallons) on a particular day. Suppose the X_i's are independent with $\mu_1 = 1000$, $\mu_2 = 500$, $\mu_3 = 300$, $\sigma_1 = 100$, $\sigma_2 = 80$, $\sigma_3 = 50$. Compute the expected value and the standard deviation of the revenue from sales

$$Y = 2.2X_1 + 2.35X_2 + 2.5X_3.$$

Bad news



In general, the mean and the variance do not define a probability distribution.

Mean and variance of the sample mean

Problem

Given a random sample $X_1, X_2, ..., X_n$ from a distribution with mean μ and standard deviation σ, the mean is modeled by a random variable $\bar{X}$,

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

- *Compute $E(\bar{X})$*
- *Compute $Var(\bar{X})$*

Mean and variance of the sample mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value μ and standard deviation σ. Then

1. $E(\overline{X}) = \mu_{\overline{X}} = \mu$

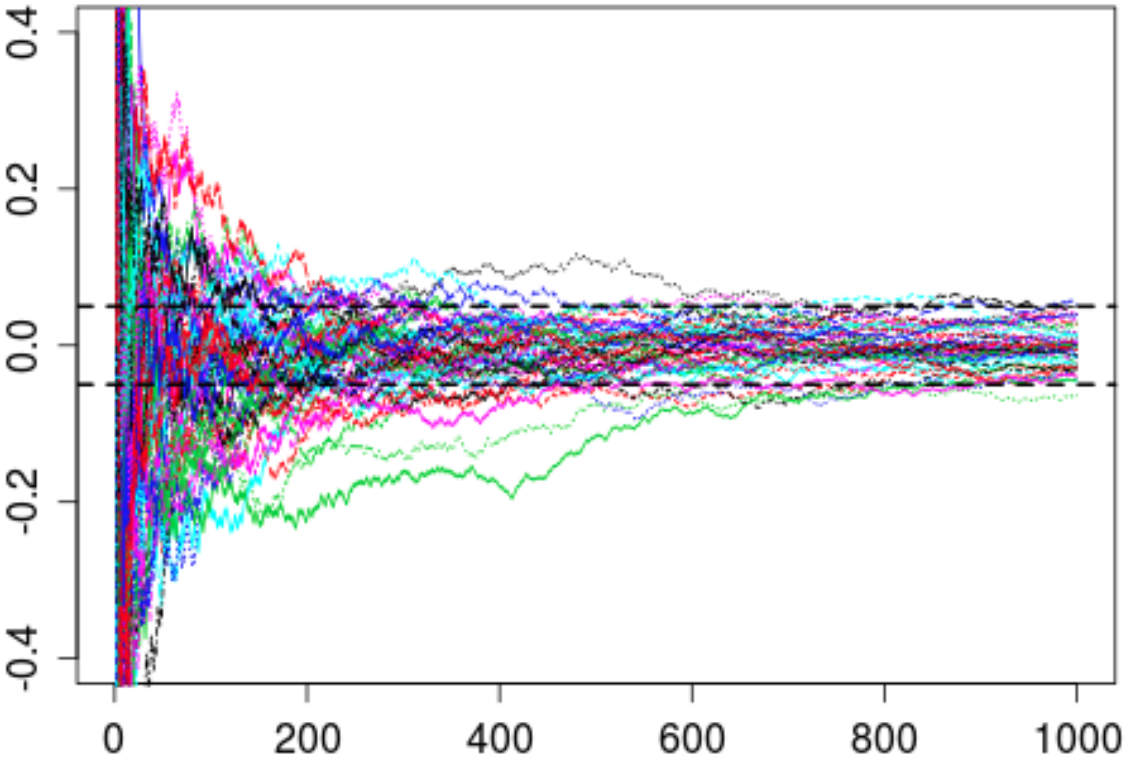
2. $V(\overline{X}) = \sigma^2_{\overline{X}} = \sigma^2/n \quad \text{and} \quad \sigma_{\overline{X}} = \sigma/\sqrt{n}$

Law of large numbers

THEOREM If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with mean μ and variance σ^2, then $\overline{X}$ converges to μ

a. In mean square $E[(\overline{X} - \mu)^2] \to 0$ as $n \to \infty$

b. In probability $P(|\overline{X} - \mu| \geq \varepsilon) \to 0$ as $n \to \infty$



The Central Limit Theorem

Theorem

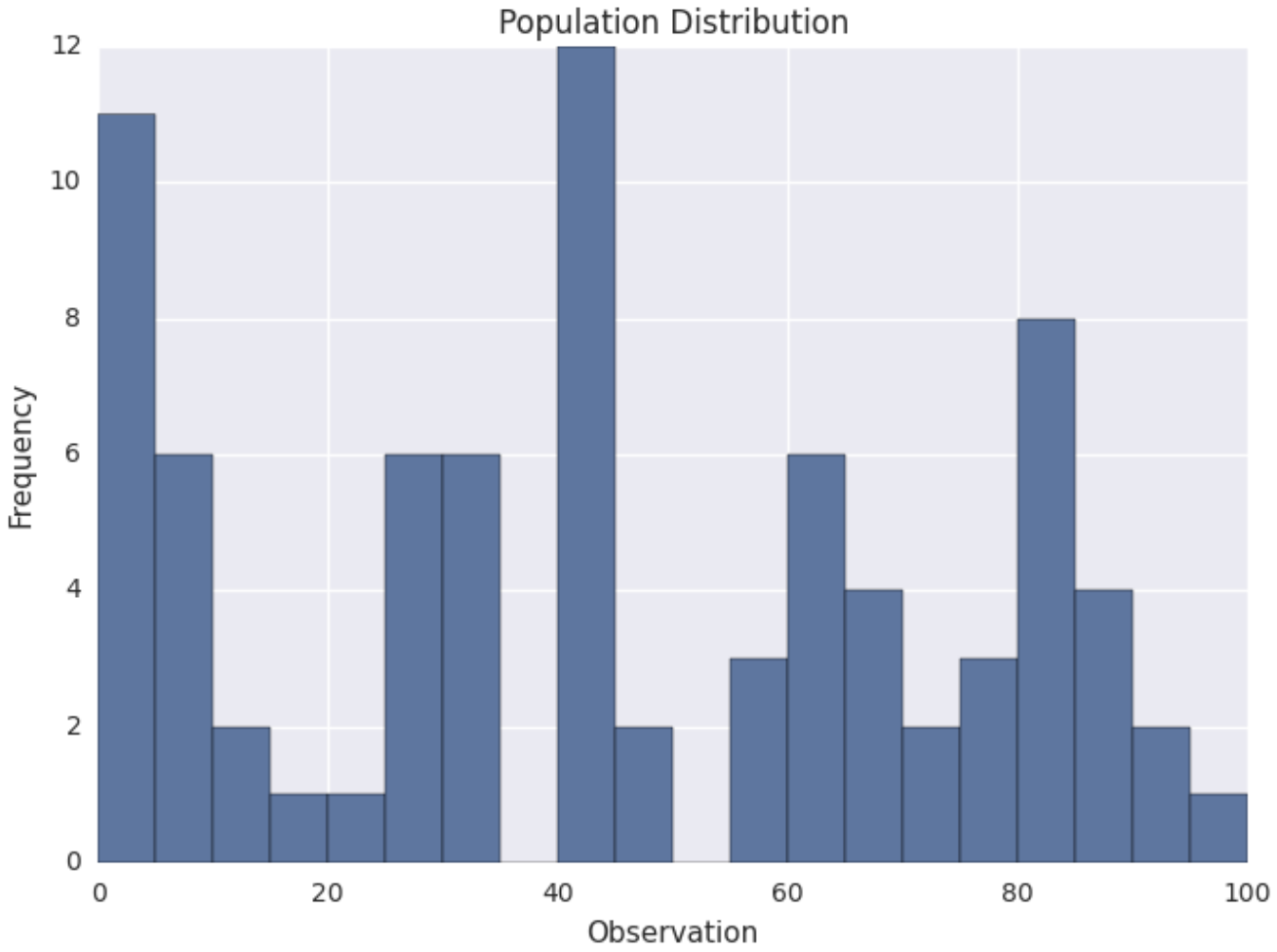
Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then, in the limit when $n \to \infty$, the standardized version of $\bar{X}$ have the standard normal distribution

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z\right) = \mathbb{P}[Z \leq z] = \Phi(z)$$

Rule of Thumb:

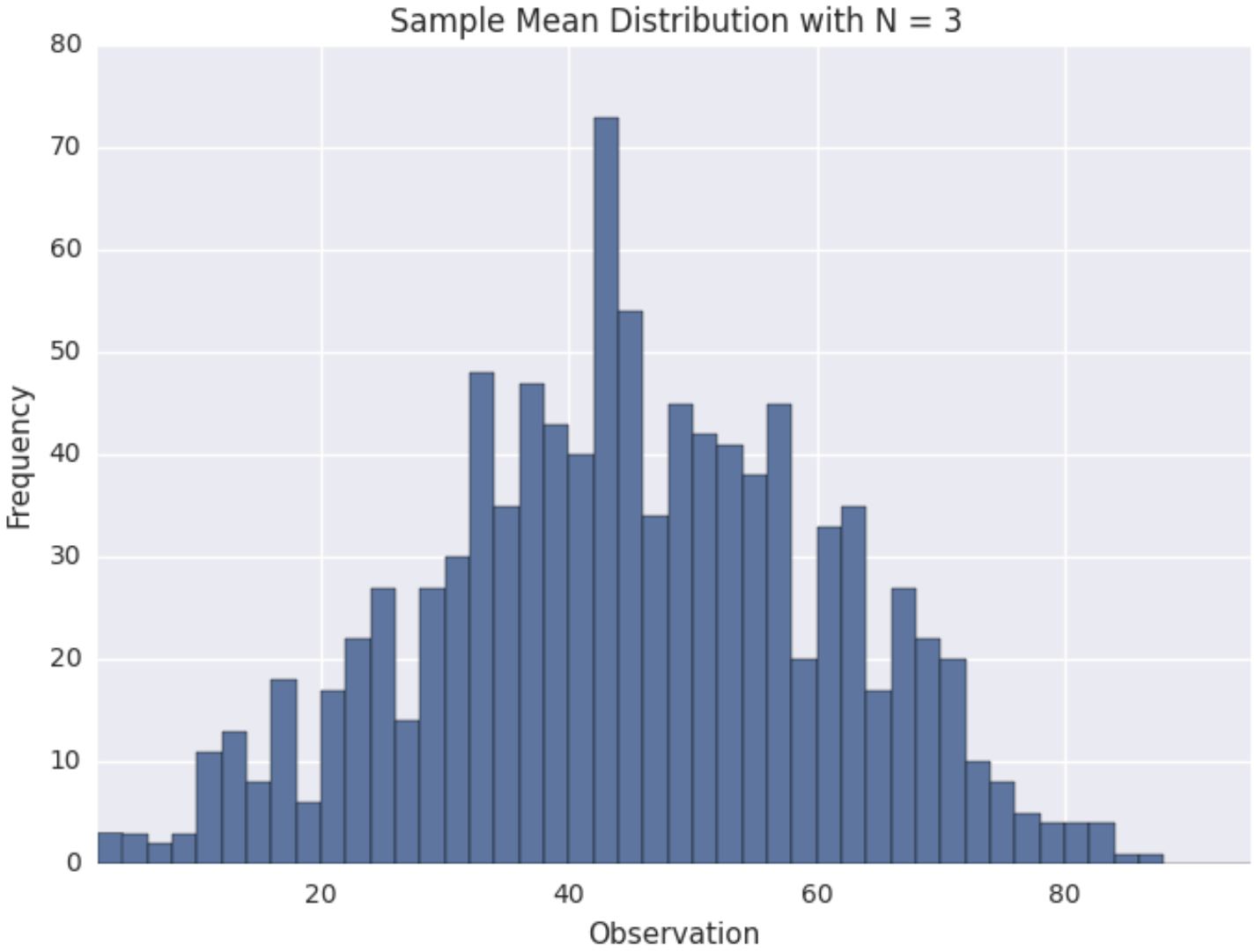
If $n > 30$, the Central Limit Theorem can be used for computation.

Example: population distribution

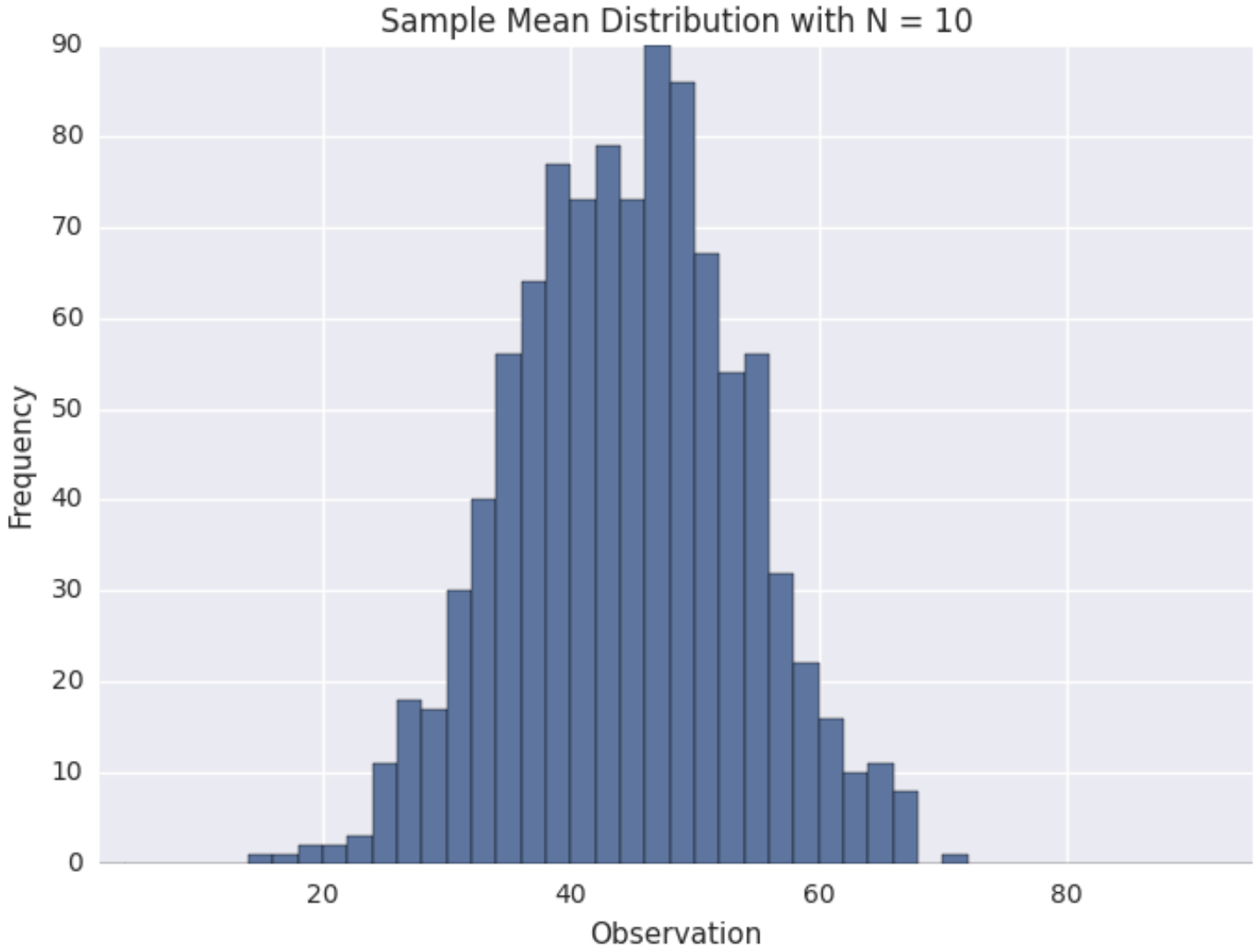


Matt Nedrick (2015).
`http://github.com/mattnedrich/CentralLimitTheoremDemo`

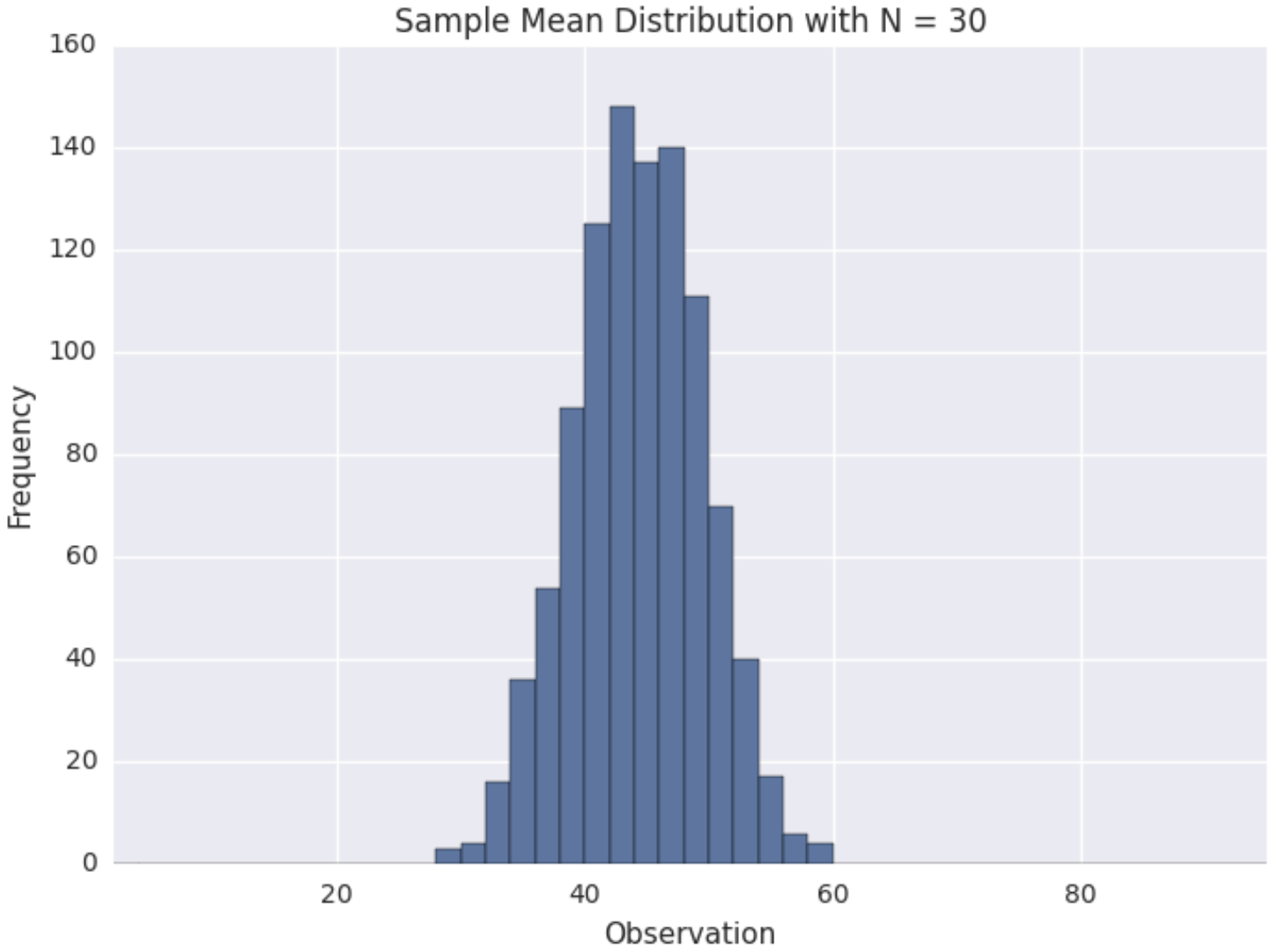
Sample distribution: $n = 3$



Sample distribution: $n = 10$



Sample distribution: $n = 30$



Example

Problem

When a batch of a certain chemical product is prepared, the amount of a particular impurity in the batch is a random variable with mean value 4.0 g and standard deviation 1.5 g.

If 50 batches are independently prepared, what is the (approximate) probability that the sample average amount of impurity X is between 3.5 and 3.8 g?

Hint:

- First, compute $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$
- Note that

$$\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

  is (approximately) standard normal.

Example

Problem

The tip percentage at a restaurant has a mean value of 18% and a standard deviation of 6%.

What is the approximate probability that the sample mean tip percentage for a random sample of 40 bills is between 16% and 19%?