Central Limit Theorem, Normal Distribution, and Inference

POSC 3410 - Quantitative Methods in Political Science

Steven V. Miller

Department of Political Science

Goal for Today

Make inferential claims from a random sample to a population.

Introduction

We are moving pretty quickly now into applied statistical inference.

- We discussed random sampling as the foundation of inference.
- This leads to an important trade-off between bias and efficiency.

We can actually calculate this random sampling error.

$$\text{R.S.E.} = \frac{\text{Variation component}}{\text{Sample size component}} \tag{1}$$

This random sampling error is the standard error of a sample mean.

$$\text{Standard error of sample mean} = \frac{\sigma}{\sqrt{n}} \tag{2}$$

What's Next?

How likely is the sample statistic given a population parameter?

- What if we assume (or even know) the population parameter?
- How likely is it we observed that sample statistic?

We can answer this question by reference to two concepts.

1. Central limit theorem
2. Normal distribution

Central Limit Theorem

The **central limit theorem** says:

- with an infinite number samples of size n...
- from a population of N units...
- the sample means will be normally distributed.

Corollary findings:

- The mean of sample means would equal μ.
- Random sampling error would equal the standard error of the sample mean ($\frac{\sigma}{\sqrt{n}}$)

Normal Distribution

A **normal distribution** is a symmetrical, continuous function.

- Its peak is the arithmetic mean (μ).
- Its width equals the variance (σ^2)

Normal Distribution

Consider Figure 6-3.

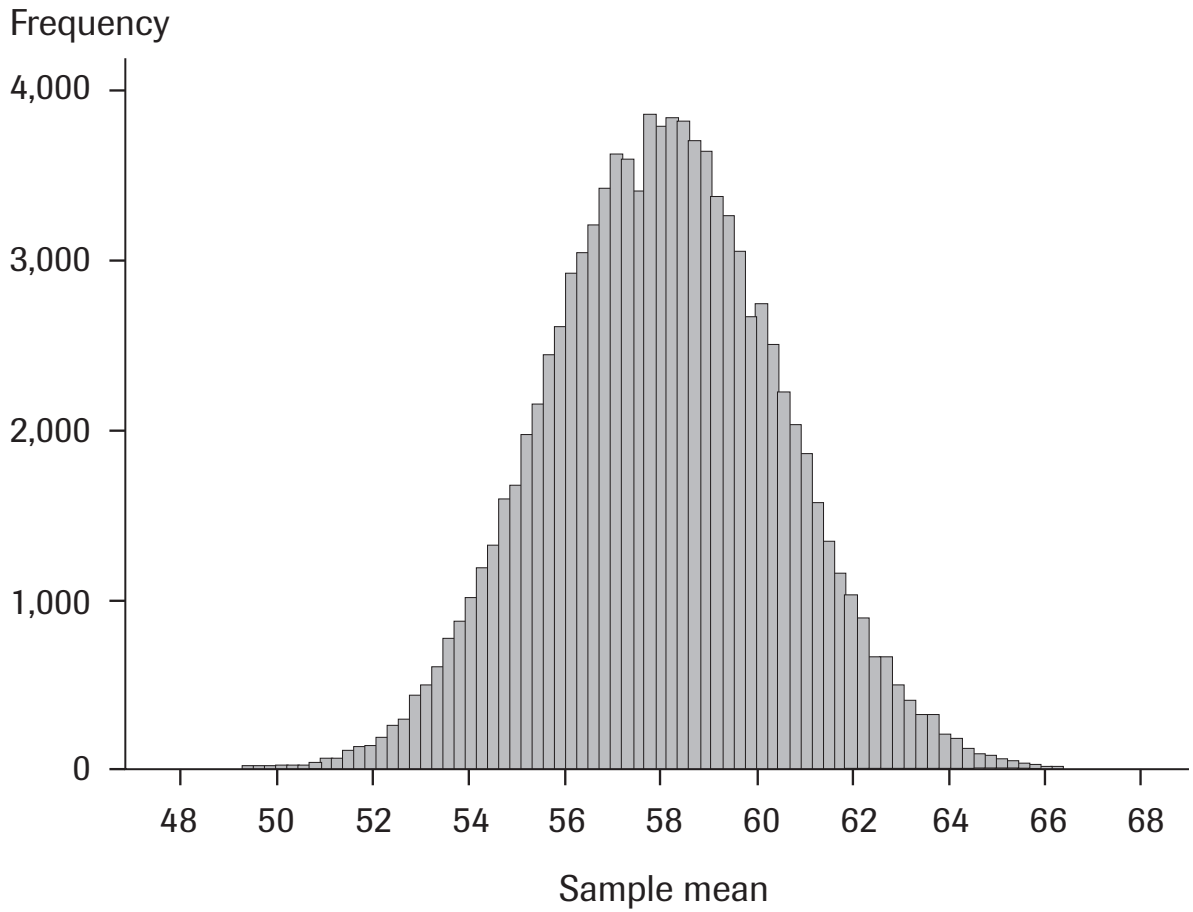
- The author has a hypothetical 20,000-student university.
- He wants to measure a "thermometer" rating of Democrats
- Assume $\mu = 58$ and $\sigma = 24.8$.

The author took 100,000 random samples of $n = 100$.

- Contrast Figure 6-3 with Panel A in Figure 6-2.

Figure 6-3

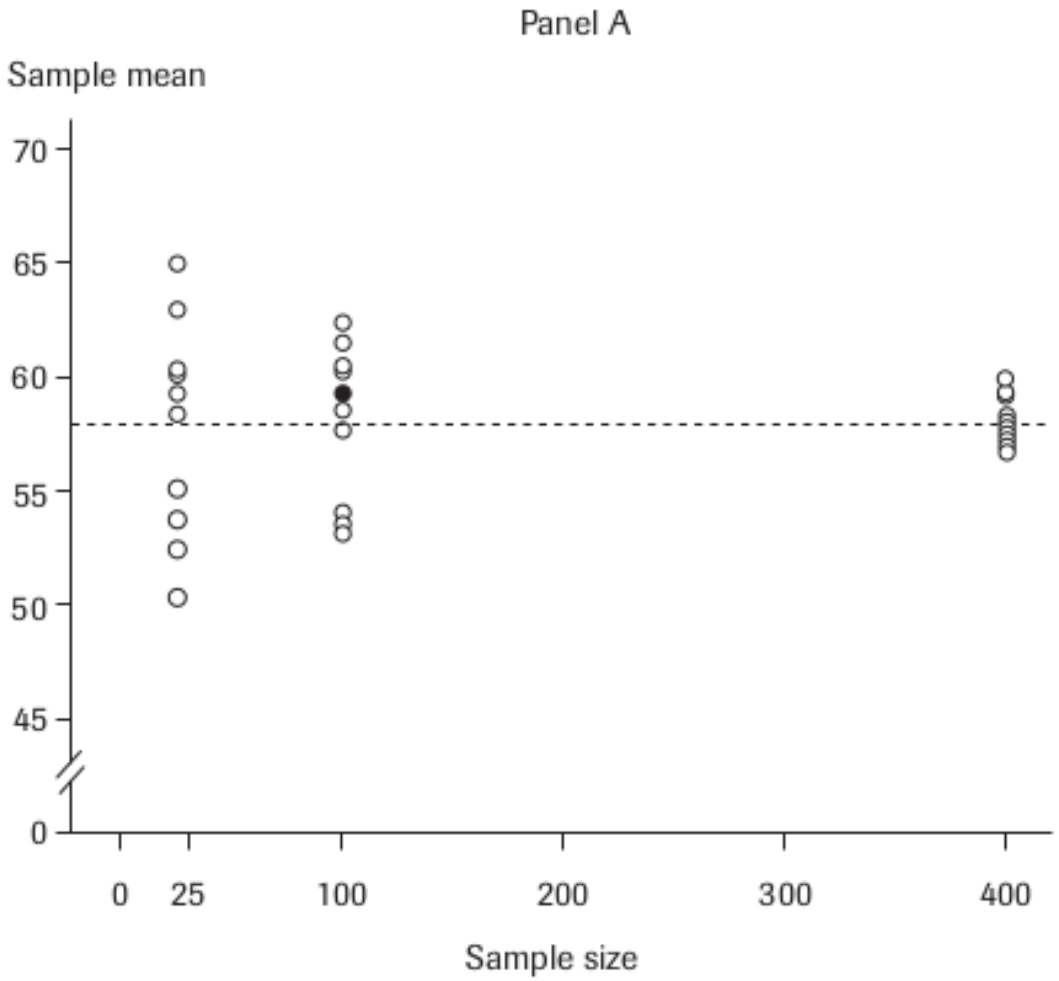
Figure 6–3 Distribution of Means from 100,000 Random Samples



Note: Displayed data are means from 100,000 samples of $n = 100$. Population parameters: $\mu = 58$ and $\sigma = 24.8$.

Figure 6-2, Panel A

Figure 6-2 Sample Means from Population with $\mu = 58$ and $\sigma = 24.8$ (Panel A) and $\sigma = 17.8$ (Panel B)



Standardization

A raw normal distribution is somewhat uninformative.

- **Standardization** will make it useful.

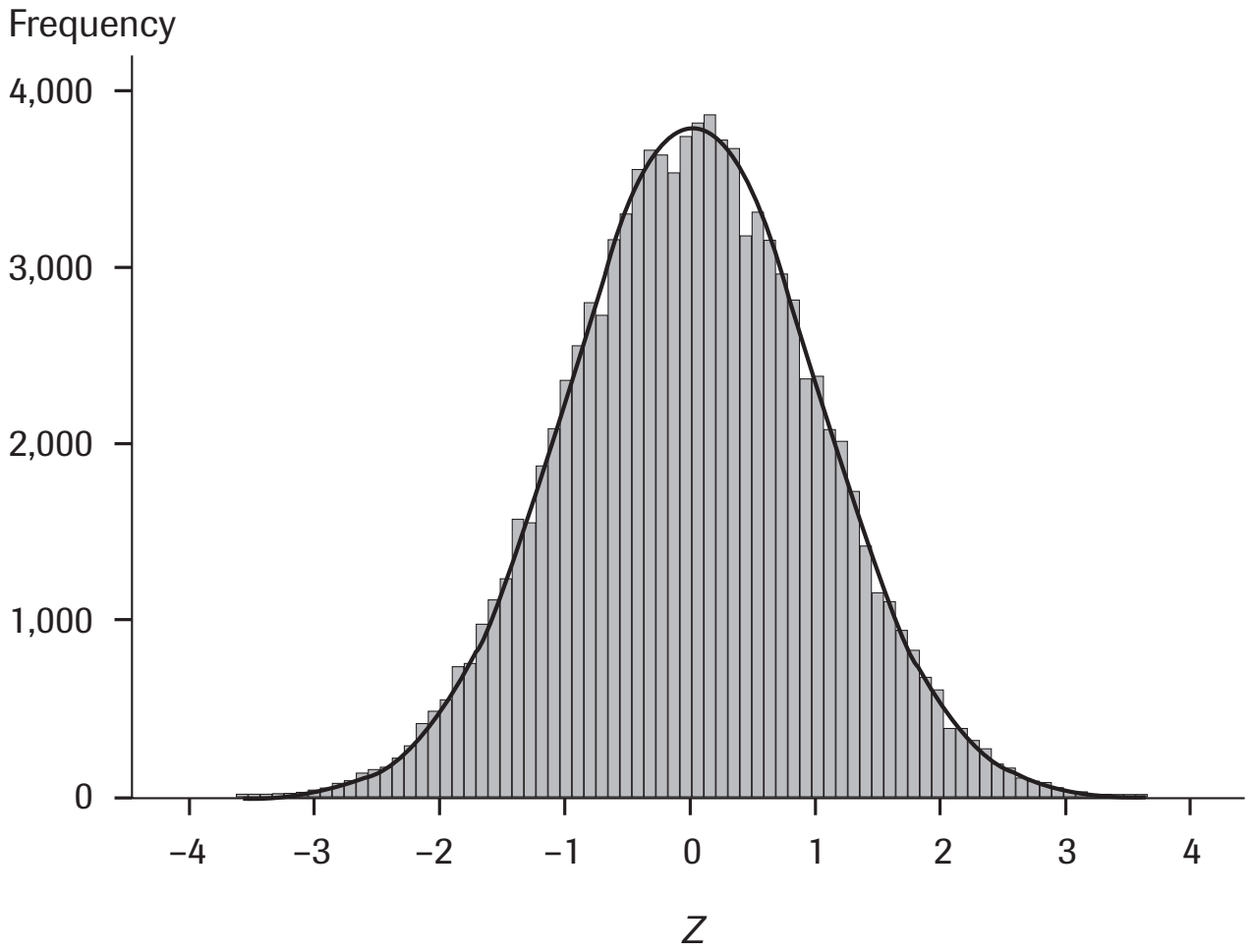
$$z = \frac{\text{Deviation from the mean}}{\text{Standard unit}} \tag{3}$$

The standard unit will vary, contingent on what you want.

- If you're working with just one random sample, it's the standard deviation.
- If you're comparing sample means across multiple random samples, it's the standard error.

Standardized Normal Distribution

Figure 6-4 Raw Values Converted to *Z* Scores



Standardization

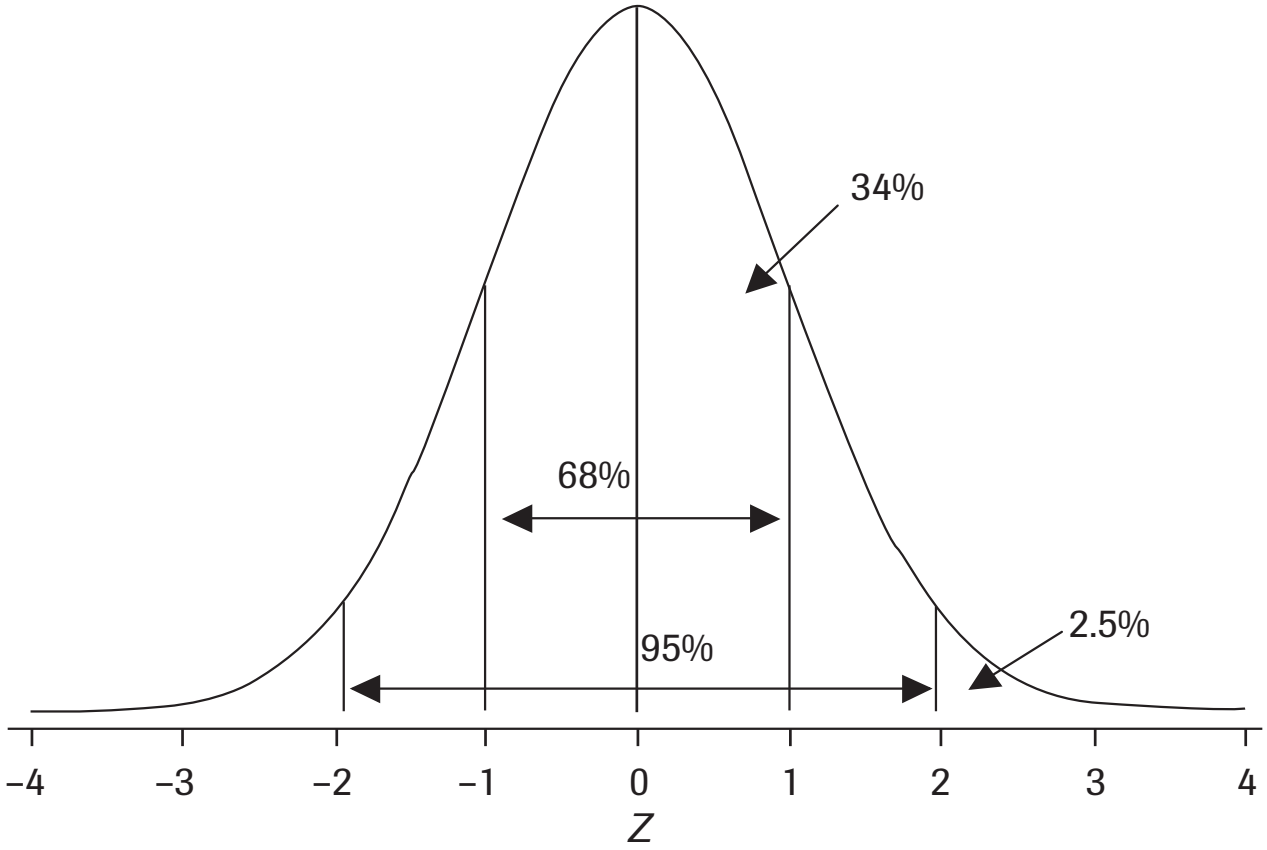
Larger z values indicate greater difference from the mean.

- When $z = 0$, there is no deviation from the mean (obviously).

Standardization allows for a better summary of a normal distribution.

Figure 6-5

Figure 6-5 Areas under the Normal Curve



Standardization and the Normal Distribution

Recall: a normal distribution is symmetrical around the peak (μ)

- Thus, 68% of cases in the distribution: $-1 \leq z \leq 1$

Notice the mark that extends to $z = |1.96|$

- This interval contains 95% of the cases in the normal distribution.

If we were to randomly pick a sample mean from the distribution, there's a 95% chance it would be within 1.96 standard errors of μ.

Inference Using the Normal Distribution

What's the next step? Assume this scenario for illustration.

- The students in our university don't know μ.
- We assume they know σ, a bit unrealistic, but alas...
- They have an n of 100 and $\overline{x}$ of 59.

They want to know the location of the population mean.

Inference Using the Normal Distribution

Our best guess of the population parameter is the sample statistic.

- We have to account for the noise introduced by random sampling.
- However, we'll never truly "know" the population parameter.

A **95-percent confidence interval** can be informative.

- It's the interval in which 95% of all possible sample estimates will fall by chance.
- We operationalize this as $\overline{x} \pm (1.96)$*(standard error).

Inference Using the Normal Distribution

How we apply this for our problem.

- We have our x-bar.
- We have our n and assume a known σ.
- Standard error $= 2.48$ ($\frac{\sigma}{\sqrt{n}} = \frac{24.8}{\sqrt{100}} = 2.48$)

We can get our upper/lower bounds of a 95-percent confidence interval.

$$\text{Lower bound} = \overline{x} - (1.96) * (2.48) = 59 - 4.8608 = 54.1392 \quad (4)$$

$$\text{Upper bound} = \overline{x} + (1.96) * (2.48) = 59 + 4.8608 = 63.8608 \quad (5)$$

Inference Using the Normal Distribution

We discuss this interval as follows.

- If we took 100 samples of $n = 100$, 95 of those random samples on average would have sample means between 54.1392 and 63.8608.

We're not saying, for the moment, the true population mean is between those two values. We don't necessarily know that.

- However, even this process gives us some nice properties.

An Illustration of Inference

Assume the College Democrats president is suspicious of our x-bar.

- (S)he claims it should be higher (say: $\mu = 66$)

So what can we do about this claim?

An Illustration of Inference

This is a probabilistic question!

- i.e. What was the probability of $\overline{x} = 59$ if $\mu = 66$?

We can answer this by reference to z values.

$$z = \frac{\overline{x} - \mu}{s.e.} = \frac{59 - 66}{2.48} = -2.82 \tag{6}$$

Find the z Value

z-Score Chart

Use this chart to find the area under a normal curve when finding an approximation for a binomial distribution.
Negative z-score - value is to the left of the mean.
Positive z-score - value is to the right of the mean.

Negative z-scores:

z	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.0
-3.4	0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005
-3.2	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007
-3.1	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010
-3.0	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013
-2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019
-2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026

An Illustration of Inference

What is the probability that a random sample would produce a z value of -2.82?

- Answer: .0024.

In other words: if μ were 66, we'd observe that $\overline{x}$ only *24 times of 10,000 samples, on average*.

- This is highly improbable.
- We suggest is the College Democrats president is likely wrong in his/her assertion.
- We offer that our sample mean is closer to what μ really is.

Some Derivations

We assumed we knew σ, if not μ. What if we don't know either?

- Use the sample standard deviation (s) instead.
- Do the same process with a **Student's t-distribution**.
- This is almost identical to a normal distribution, but with fatter tails for fewer **degrees of freedom**.

Uncertainty increases with fewer degrees of freedom.

Student's t-distribution

Table of Probabilities for Student's t-Distribution

df	0.600	0.700	0.800	0.900	0.950	0.975	0.990	0.995
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660
120	0.254	0.526	0.845	1.289	1.658	1.980	2.358	2.617

df (degrees of freedom) = number of samples - 1
1 - alpha (for one tail) or 1 - alpha/2 (for two tails)

Some Derivations

What about **sample proportions**? Let p = proportion of cases in one category.

$$\text{Standard error of sample proportion} = \frac{\sqrt{p * (1 - p)}}{\sqrt{n}} \tag{7}$$

From there, do the same process you've done previously with z values.

- *Important*: inference is unreliable when p is very small ($p < .05$).

Conclusion: The Process of Inference

Notice the process of inference.

1. Assume the hypothetical mean to be correct.
2. Test the claim about the hypothetical mean based on a random sample.
3. Infer about the claim of the population mean using probabilistic inference.

We will never know μ, but we know more about μ by randomly sampling the population and determining what μ is likely not.

Table of Contents