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.
- While unfortunate, we can calculate 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 a 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 through reference to two concepts.

1. Central limit theorem
2. Normal distribution

Central Limit Theorem

The **central limit theorem** says an infinite number of samples of size n from a population of N units will have sample means that are normally distributed.

The mean would equal the population mean (μ) and have a random sampling error equal to the standard error of the sample mean ($\sigma/\sqrt{n}$).

Normal Distribution

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

- 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

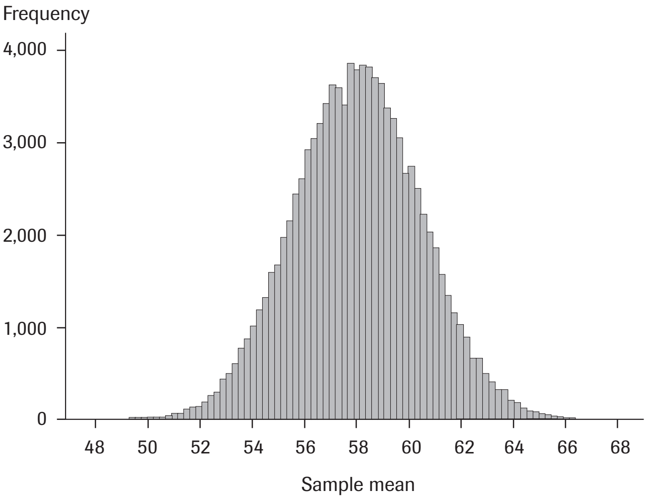
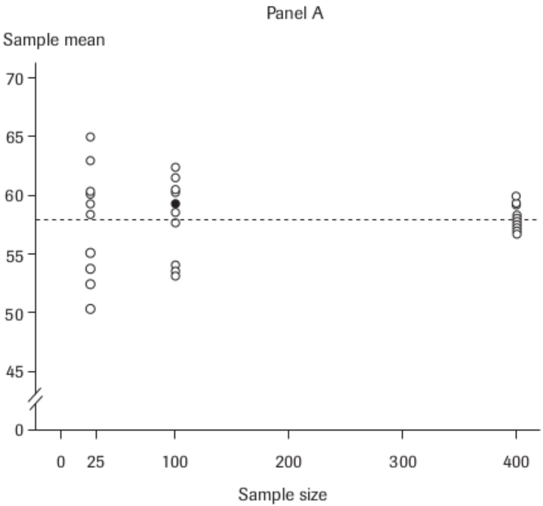


Figure 6-2, Panel A

Figure 6-2 Sample Means from Population with μ = 58 and σ = 24.8 (Panel A) and σ = 17.8 (Panel B)



Standardization

A transformation of a normal distribution proves very handy.

We do this through **standardization**, in which we divide the deviation of a value from the mean over a standard unit.

- This gives us a Z value.

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

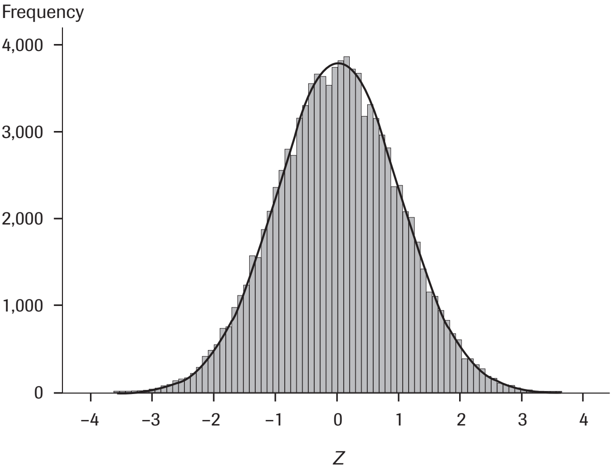
The standard unit will vary, contingent on what you want.

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

Standardization

When you standardize raw values, you get the following distribution.

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



Standardization

As you might have guessed, larger Z values indicate greater difference from the mean.

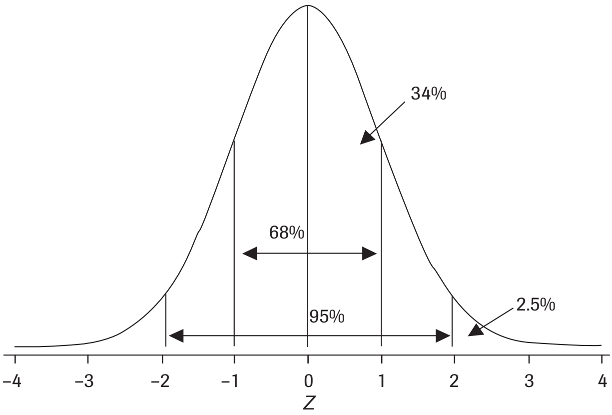
- When Z = 0, there is no deviation from the mean. It is the mean.

Why do this? Why bother?

- Standardization allows for a better summary of a normal distribution.
- Consider Figure 6-5.

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, we can say 68% of cases in the distribution have a Z score between -1 and 1.
- Notice the mark extending to $Z = |1.96|$.
  - This contains 95% of cases in the normal distribution.

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

- *Keep this in mind going forward.*

Inference Using the Normal Distribution

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

- The students in our hypothetical university don't know μ.
- We assume they know σ, a bit unrealistic, but alas.
- They have a n of 100 and a $\overline{x}$ 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 noise introduced by random sampling.
- However, we'll never truly "know" the population parameter.

The standard of a **95 percent confidence interval** helps.

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

Inference Using the Normal Distribution

How do we apply this to our previous example?

- We have our x-bar ($\overline{x} = 59$).
- We have our n ($n = 100$) and assume a known σ ($\sigma = 24.8$).
- Standard error: 2.48. ($\frac{\sigma}{\sqrt{n}} = \frac{24.8}{\sqrt{100}} = 2.48$)

Let's get our upper and lower bounds of a 95 percent confidence interval.

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

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

Inference Using the Normal Distribution

We discuss this 95 percent confidence interval as follows.

- If we took 100 samples of $n = 100$, 95 of those random samples would have sample means between 54.1392 and 63.8608.
- We are <u>not</u> saying the *true* population mean is between those two values. We don't necessarily know that.

An Illustration of Inference

However, even this process gives us nice properties.

Assume the College Democrats president is suspicious of that sample mean.

- (S)he believes it is higher and that $\mu = 66$.

What can we say about this claim?

An Illustration of Inference

This is an issue of **probability**, the likelihood of an event occurring.

- Basically, what was the probability of $\overline{x} = 59$ if $\mu = 66$?

We can answer this by reference to Z statistics!

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

This Z score will allow us to calculate the statistical distance between the proposed population mean and the sample statistic.

Table 6-3 Proportions of the Normal Curve above the Absolute Value of *Z*

First digit and first decimal of *Z*	Second decimal of *Z*									
	.00	*.01*	*.02*	*.03*	*.04*	*.05*	*.06*	*.07*	*.08*	*.09*
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	**.0250**	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

An Illustration of Inference

What is the percentage of possible random samples that would produce a Z score of -2.82?

- Answer: .0024

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

- This is highly improbable.
- We suggest the College Democrats president is very likely wrong in his/her assertion.

Some Derivations

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

- Use the sample standard deviation (s) instead of σ.
- 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**, which are the number of observations minus the number of parameters being estimated.

Table 6-4 The Student's *t*-Distribution

Degrees of freedom	Area under the curve			
	.10	*.05*	*.025*	*.01*
1	3.078	6.314	12.706	31.821
2	1.886	2.920	4.303	6.965
3	1.638	2.353	3.182	4.541
4	1.533	2.132	2.776	3.747
5	1.476	2.015	2.571	3.365
6	1.440	1.943	2.447	3.143
7	1.415	1.895	2.365	2.998
8	1.397	1.860	2.306	2.896
9	1.383	1.833	2.262	2.821
10	1.372	1.812	2.228	2.764
11	1.363	1.796	2.201	2.718
12	1.356	1.782	2.179	2.681
13	1.350	1.771	2.160	2.650
14	1.345	1.761	2.145	2.624
15	1.341	1.753	2.131	2.602
16	1.337	1.746	2.120	2.583
17	1.333	1.740	2.110	2.567
18	1.330	1.734	2.101	2.552
19	1.328	1.729	2.093	2.539
20	1.325	1.725	2.086	2.528
21	1.323	1.721	2.080	2.518
22	1.321	1.717	2.074	2.508
23	1.319	1.714	2.069	2.500
24	1.318	1.711	2.064	2.492
25	1.316	1.708	2.060	2.485
26	1.315	1.706	2.056	2.479
27	1.314	1.703	2.052	2.473
28	1.313	1.701	2.048	2.467
29	1.311	1.699	2.045	2.462
30	1.310	1.697	2.042	2.457
40	1.303	1.684	2.021	2.423
60	1.296	1.671	2.000	2.390
90	1.291	1.662	**1.987**	2.368
100	1.290	1.660	1.984	2.364
120	1.289	1.658	1.980	2.358
1,000	1.282	1.646	1.962	2.330
Normal (Z)	1.282	1.645	1.960	2.326

Some Derivations

What if we're dealing with **sample proportions**, the number of cases in one category divided over the total number of observations?

Assume $p =$ 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.

- 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 can know more about μ by randomly sampling the population and determining what μ is not.

Table of Contents