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?

1. What if we assume (or even know) the population parameter?
2. 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)

You should remember some other features our lecture on this distribution.

An Applied Example from a Thermometer Rating

Let’s use a real-world illustration from the 2020 ANES exploratory testing survey.

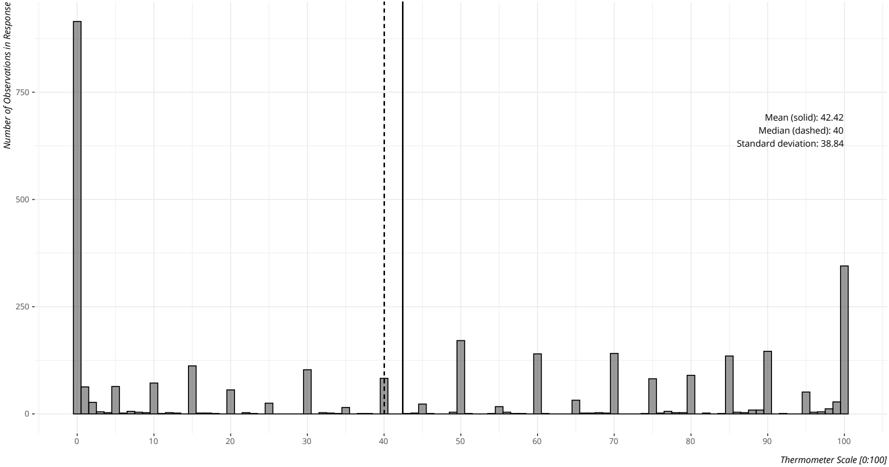
- Survey period: April 10-18, 2020 (online).
- Released July 27, 2020

The question is a basic thermometer rating of Donald Trump.

- Scale: 0 (“coldest”) to 100 (“warmest”)

Thermometer Ratings for Donald Trump (ANES ETS, 2020)

Thermometer ratings for divisive political figures in the U.S. tend to be ugly as hell with estimates of central tendency that don't faithfully capture the data.



Data: American National Election Studies (Exploratory Testing Survey, 2020). N = 3,073.

Thermometer Rating

This is what you'll get in these questions, by the way.

- 0-100 thermometer ratings are *noisy* with natural "heaping" patterns.
- Partisanship only compounds the problem, as you can probably gather.

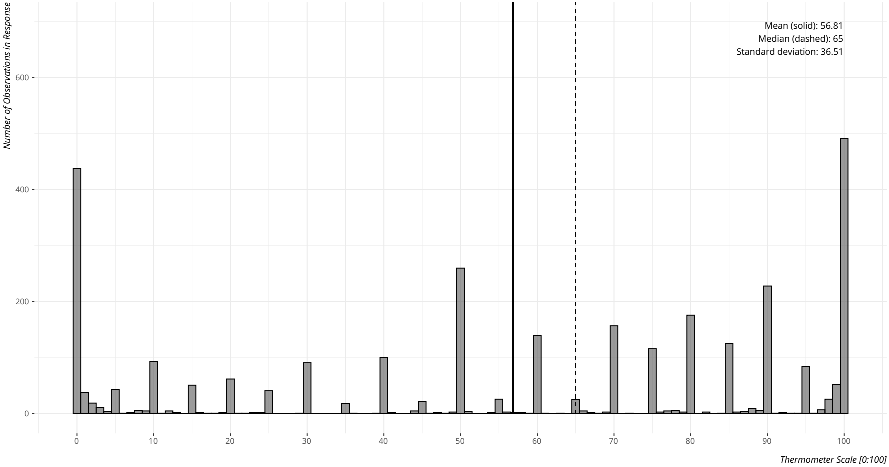
Notice the mean and median aren't too far apart, but the data don't look "normal" for obvious reasons.

- Standard deviation is also huge.
- Conventional interpretation: there are more people who loathe Trump than those who really love him.

Next slide: what it looks like for Barack Obama.

Thermometer Ratings for Barack Obama (ANES ETS, 2020)

Again: thermometer ratings for divisive political figures in the U.S. tend to be ugly as hell with estimates of central tendency that don't faithfully capture the data.



Data: American National Election Studies (Exploratory Testing Survey, 2020). N = 3,072.

What We'll Do

Let's create a hypothetical "population" with the set parameters from the Trump ratings.

- Data will be bound between 0 and 100 with a mean of 42.42 and standard deviation of 38.84.
- N = 250,000 (i.e. scaled down from U.S. adult population of ~250 million).

We want to approximate the "population" mean thermometer rating via central limit theorem.

- We'll grab a million samples of ten respondents and store the sample means.

Let's plot the results.

R Code

```r
# rbnorm from stevemisc
Population <- rbnorm(250000, mean =42.42, sd = 38.84,
                     lowerbound = 0,
                     upperbound = 100,
                     round = TRUE,
                     seed = 8675309) # Jenny, I got your number...
```

Note: it’s hard to perfectly mimic these kind of thermometer ratings from a simple distribution, but this will do.

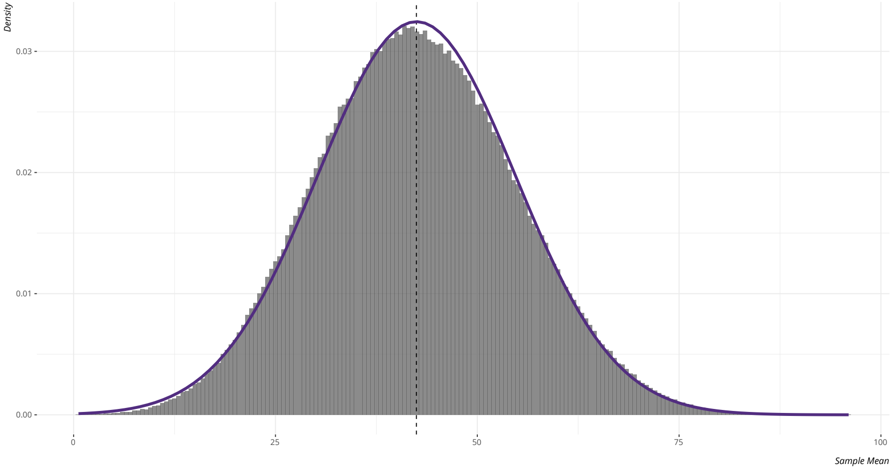
- Mean: 42.459772
- Standard deviation: 38.8881803

R Code

```r
set.seed(8675309) # Jenny, I got your number...
# Note dqrng offers much faster sampling at scale
Popsamples <- tibble(
  samplemean=sapply(1:1000000,
           function(i){ x <- mean(
             dqsample(Population, 10,
                    replace = FALSE))
           }))
```

The Distribution of 1,000,000 Sample Means, Each of Size 10

Notice the distribution is normal and the mean of sample means converges on the known population mean (vertical line).



Data: Simulated data for a population of 250,000 where mean = 42.42 and standard deviation = 38.84.

How Did We Do?

See for yourself:

```
mean(Popsamples$samplemean)
```

```
## [1] 42.47174
```

```
mean(Population)
```

```
## [1] 42.45977
```

Not bad...

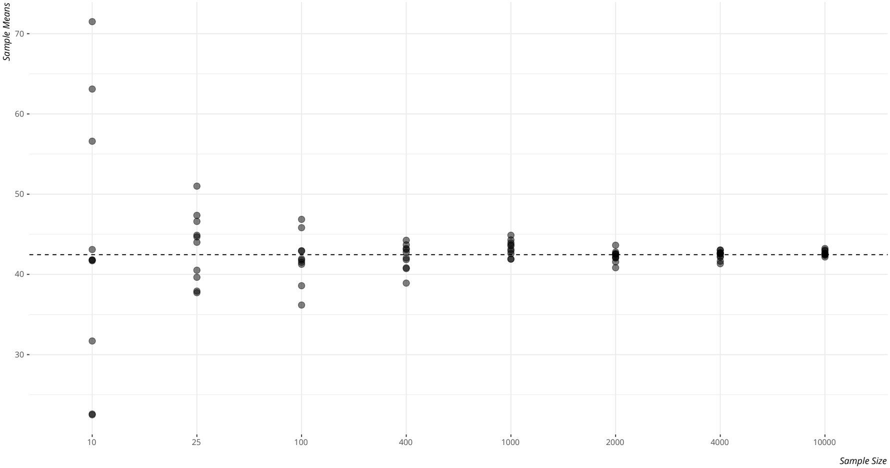
Implications of Central Limit Theorem

Remember the previous implications of random sampling error?

- i.e. a good-sized sample reduces random sampling error in even high-variation data?

Ten Sample Means of Varying Sample Sizes from a Population

The diminishing returns of increasing sample size emerge around 1,000 observations, even as the spread in these simulated data is quite large.



Data: Simulated data for a population of 250,000 where mean = 42.42 and standard deviation = 38.84.

Implications of Central Limit Theorem

Likewise, infinite samples of *any* size (even absurdly small samples of high-variation data) reduce the gap between estimate and “true” population parameter.

Standardization

A raw normal distribution I presented 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.

Standardization

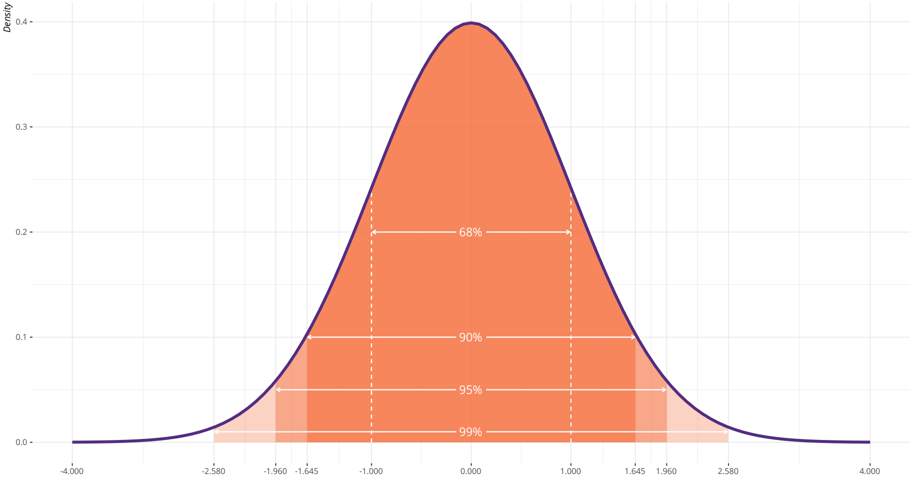
Larger z values indicate greater difference from the mean.

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

Standardization allows for a better summary of a normal distribution.

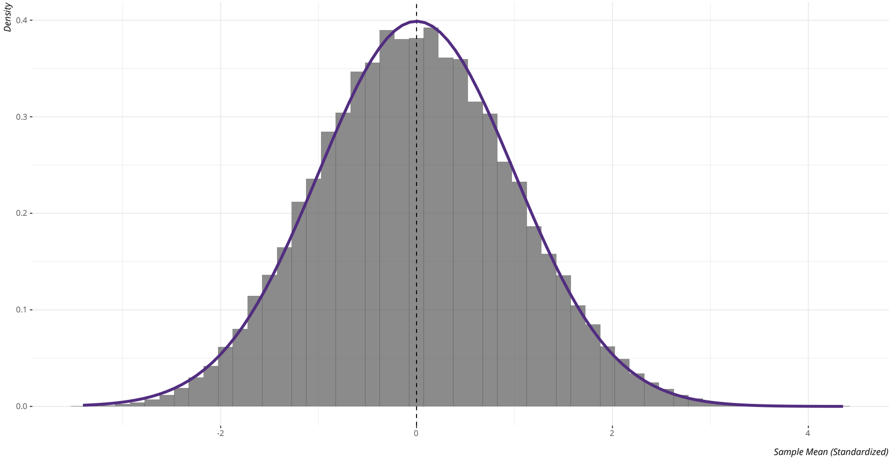
The Area Underneath a Normal Distribution

The tails extend to infinity and are asymptote to zero, but the full domain sums to 1. 95% of all possible values are within about 1.96 standard units from the mean.



The Distribution of 1,000,000 Sample Means, Each of Size 10

Notice the distribution is normal and the mean of sample means converges on the known population mean (vertical line).



Data: Simulated data for a population of 250,000 where mean = 42.42 and standard deviation = 38.84.
Histograms are admittedly a bit clunky here because the choice of bin width may be misleading, but this is at least honest.

Inference Using the Normal Distribution

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

- We as researchers have a sample of 100 people from this population.

```
set.seed(8675309)
oursample <- sample(Population, 100, replace = FALSE)
mean(oursample)
```

```
## [1] 43.64
```

- We as researchers don't know μ (though it's 42.46).
- We assume we know σ (38.89), a bit unrealistic, but alas...
- We have an *n* of 100 and $\overline{x}$ of 43.64.

We want to know the location of the population mean.

Inference Using the Normal Distribution

Our best guess of the population parameter from the sample 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 = 3.889 ($\frac{\sigma}{\sqrt{n}} = \frac{38.88}{\sqrt{100}} = 3.88$)

Inference Using the Normal Distribution

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

$$\text{Lower bound} = \overline{x} - (1.96) * (s.e.) \tag{4}$$

$$\text{Upper bound} = \overline{x} + (1.96) * (s.e.) \tag{5}$$

R Code

```r
#computation of the standard error of the mean
sem<-sd(Population)/sqrt(length(oursample))
#95% confidence intervals of the mean
c(mean(oursample)-1.96*sem, mean(oursample)+1.96*sem)
```

```
## [1] 36.01792 51.26208
```

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 36.02 and 51.26.

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 we have a Pickens County resident who is suspicious of our *x*-bar.

- (S)he claims it has to be much higher. Say: 56.61.
  - Rationale: this is the percentage of the vote Trump got in the Pike precinct of Pickens County.
  - In other words, (s)he is basically inferring by anecdote or making hasty generalizations from his/her surroundings.

So what can we do about this claim?

An Illustration of Inference

This is a probabilistic question!

- i.e. What was the probability of $\overline{x}$ = 43.64 if μ = 56.61?

We can answer this by reference to *z* values.

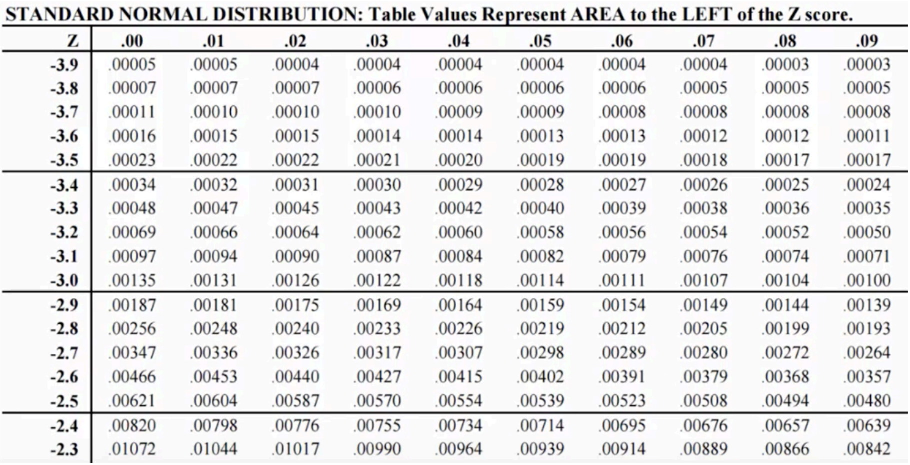
$$z = \frac{\overline{x} - \mu}{s.e.} \tag{6}$$

R Code

```r
(mean(oursample)-56.61)/sem
```

```
## [1] -3.335204
```

Find the *z* Value



STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842

...or in R

```r
1-pnorm(abs((mean(oursample)-56.61)/sem))
```

```
## [1] 0.0004261848
```

An Illustration of Inference

What is the probability that a random sample would produce a *z* value of -3.3352?

- Answer: 0.00043

In other words: if μ were 56.61, we'd observe that $\overline{x}$ only about 4 times in 10,000 trials, on average.

- This is highly improbable.

An Illustration of Inference

What do we conclude?

- We suggest this hypothetical Pickens County resident is likely wrong in his/her assertion.
- We offer that our sample mean is closer to what μ really is.

Since we've been playing god this whole time, we know that's true.

What About the Known Population Mean?

How likely was our $\overline{x}$ of 43.64 given the μ of 42.46? Same process.

```
(mean(oursample)-mean(Population))/sem
```

```
## [1] 0.3034927
```

```
1-pnorm(abs((mean(oursample)-mean(Population))/sem))
```

```
## [1] 0.3807572
```

The probability of our sample mean, given the population mean (that we know), is 0.38.

- This is a likely outcome.
- We cannot rule out the population mean from our random sample like we could with the hypothetical mean of 56.61.

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**.
    - Degrees of freedom = n - k (i.e. number of observations - number of parameters [here: 1])

Uncertainty increases with fewer degrees of freedom.

Student’s t-distribution

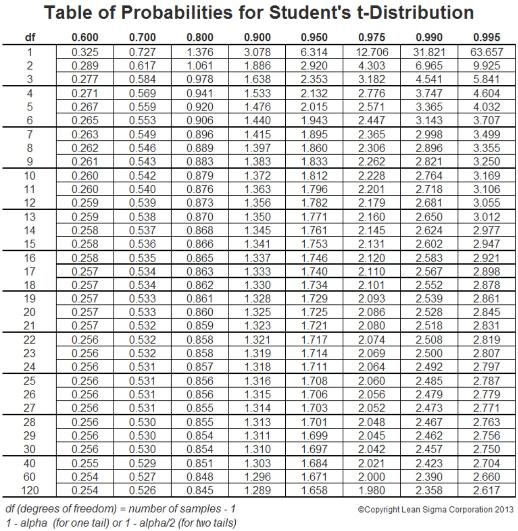


Table of Probabilities for Student's t-Distribution

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

Introduction

Central Limit Theorem and Normal Distribution

- Standardization
- Inference Using the Normal Distribution

Conclusion