Hypothesis Testing

EH6105 – Quantitative Methods

Steven V. Miller

Department of Economic History and International Relations

Goal(s) for Today

1. Revisit the logic of (infinity) random sampling from a knonw population.
2. Introduce students to the so-called “confidence” interval.
3. Reiterate what sample inference to a population actually is.

A Brief Aside

Today is going to be "hypothesis testing" vis-a-vis a sample and the population.

- i.e. what is the probability of the sample statistic, given the population parameter?

What it's not, but you should know anyway: framing hypotheses from your theories.

What is a Hypothesis (by Way of Theory)?

Hypotheses are testable statements about a relationship between an independent variable and a dependent variable.

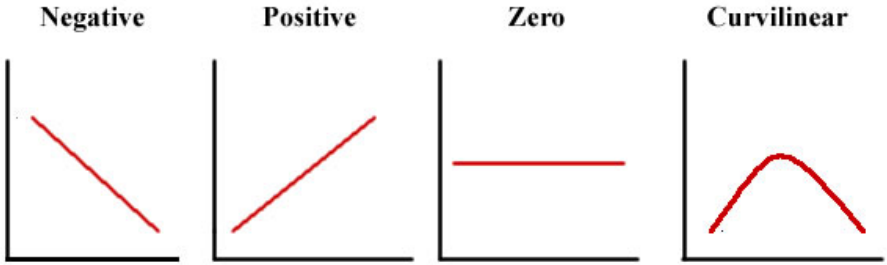
- Dependent variable: the thing you want to explain.
- Independent variable: the thing you believe explains variation in the dependent variable.

What Should Hypotheses Say?

Hypotheses must communicate the following:

1. A clear identification of proposed cause and effect
2. The proposed relationship expected between both variables
3. The unit of analysis
4. An unambiguous indication of the type of measurement in both variables.

Types of Proposed Relationships



Making Guesses About the Population

Let's revisit the previous case we used (inspired by American attitudes re: Trump).

- We have a hypothetical politician who is more despised than revered.
- Population (n = 250,000) is evaluating the politician with a thermometer rating [0:100]
- We, the gods creating Population, assign known population parameters.

We want to make guesses about Population based on samples of Population.

Creating the Data

Let's revisit the data we created.

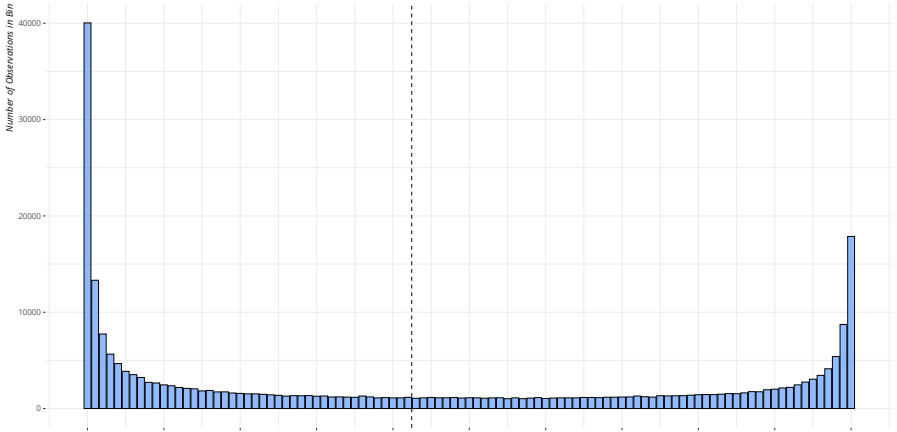
```
# rbnorm() from {stevemisc}
Population <- rbnorm(250000, mean = 42.42, sd = 38.84,
                     lowerbound = 0,
                     upperbound = 100,
                     round = TRUE,
                     seed = 8675309) # Jenny, I got your number...
```

And the summary statistics.

```
mean(Population)
#> [1] 42.45977
sd(Population)
#> [1] 38.88818
# ^ these are the gospel truth about Population
```

The Distribution of Thermometer Ratings from our Population

These data approximate the shape and distribution of real-world thermometer ratings of divisive public officials.



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

Central Limit Theorem

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:

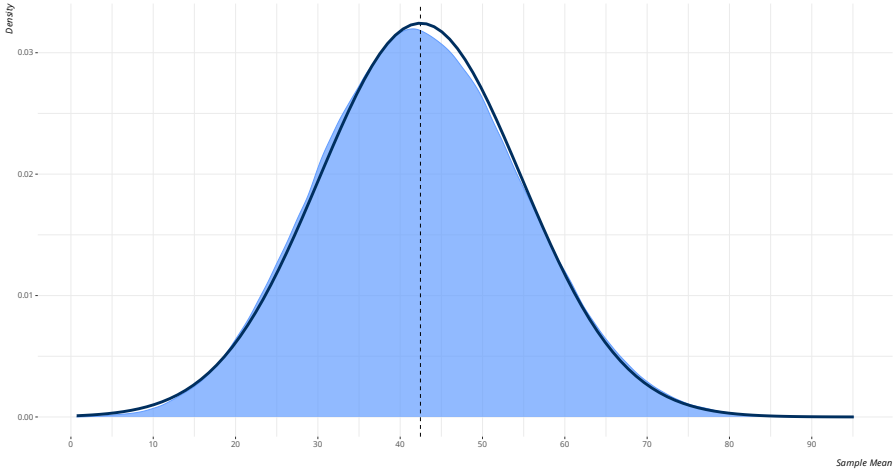
- The mean of sample means would equal μ.
- Random sampling error would equal the standard error of the sample mean ($\frac{\sigma}{\sqrt{n}}$).

R Code

```
set.seed(8675309) # Jenny, I got your number...
# Note {dqrng} offers much faster sampling at scale
# This is the dqsample() function
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.

Standardization

A raw normal distribution I presented is somewhat uninformative.

- **Standardization** will make it useful.

$$z = \frac{\text{Deviation from the mean}}{\text{Standard unit}} \tag{1}$$

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

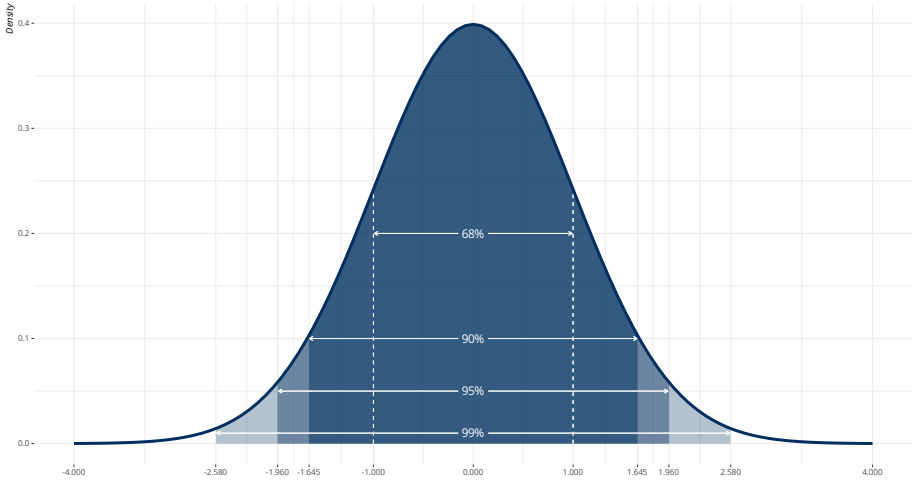
Larger z values indicate greater difference from the mean.

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

Standardization allows for a better summary of a normal distribution.

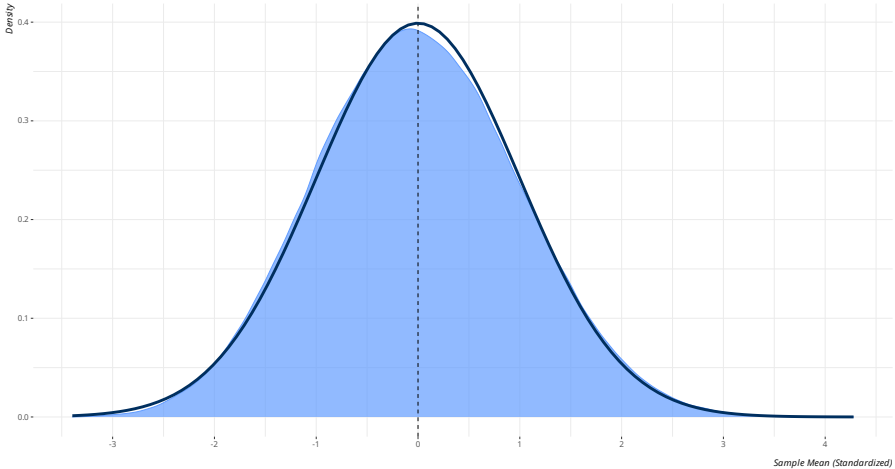
The Area Underneath a Normal Distribution

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.

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 make a statement about 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 $\overline{x}$.
- 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.

$$\text{Lower bound} = \overline{x} - (1.96) * (s.e.) \tag{2}$$

$$\text{Upper bound} = \overline{x} + (1.96) * (s.e.) \tag{3}$$

A Brief Aside...

If we're going to do inference the wrong way, we should at least get the z-values right.

```
# p_z() is in {stevemisc}

p_z(.32) # this is not 1
#> [1] 0.9944579
p_z(.10) # this is not 1.645
#> [1] 1.644854
p_z(.05) # this is not 1.96
#> [1] 1.959964
p_z(.01) # this is not 2.58
#> [1] 2.575829
p_z(0) # okay, that's still infinity
#> [1] Inf
```

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 politician's supporter who is suspicious of our $\overline{x}$.

- They claims it has to be much higher. Say: 56.61.
  - Rationale: this is the percentage of the vote Trump in the precinct where I lived during the 2020 election.
  - In other words, they are 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.

$$z = \frac{\overline{x} - \mu}{s.e.} \tag{4}$$

R Code

```
(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

Figure 1: Find the *z* value

...or in R

```r
# one-tailed (i.e. I'm assuming the direction)
pnorm(abs((mean(oursample) - 56.61)/sem), lower.tail=FALSE)
#> [1] 0.0004261848
# two-tailed (i.e. I don't know the direction)
# Notice there really isn't much happening in this distinction
# "two-tail" is a sort of default, but it's kind of silly that it is.
2*pnorm(abs((mean(oursample) - 56.61)/sem), lower.tail=FALSE)
#> [1] 0.0008523696
```

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 supporter is likely wrong in their assertion.
- We offer that our sample mean is closer to what μ really is.

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

- However, this procedure doesn't necessarily tell you what μ is.
- It's communicating what you think it's highly unlikely to be.

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
# One tail (i.e. I'm assuming the direction)
pnorm(abs((mean(oursample) - mean(Population))/sem),
        lower.tail=FALSE)
#> [1] 0.3807572
# Two tail (i.e. I'm agnostic about the direction)
2*pnorm(abs((mean(oursample) - mean(Population))/sem),
        lower.tail=FALSE)
#> [1] 0.7615144
```

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)

Figure 2: Student's t-distribution

...or in R

```r
# proposed mean
(mean(oursample) - 56.61)/
  (sd(oursample)/sqrt(100))  -> tstat1
 # actual mean
(mean(oursample) - mean(Population))/
  (sd(oursample)/sqrt(100)) -> tstat2

# probability of what we got, if the politician's supporter is right
pt(-abs(tstat1), df = 100-1) # one tail
#> [1] 0.0006489941
2*pt(-abs(tstat1), df = 100-1) # two tail
#> [1] 0.001297988
# probability of what we got, knowing what the population mean is
pt(-abs(tstat2), df = 100-1) # one tail
#> [1] 0.3819116
2*pt(-abs(tstat2), df = 100-1) # two tail
#> [1] 0.7638231
```

Conclusion: The Process of Inference

Notice the process of inference.

1. Assume the hypothetical mean to be correct (as a hypothesis, if you will).
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.

Does that look familiar? It's $p(\bar{x}|\mu)$.

- Notice what it's *not*? $p(\mu|\bar{x})$.
- That's not the question you're asking but it's the answer you're getting.

Conclusion: The Process of Inference

We will never know μ.

- But this process gives an indirect answer to the question you're asking.
- *Within* a desired confidence interval: "I can't rule these out."
- *Outside* a desired confidence interval: "what I got is highly unlikely if what you're proposing were actually true."

Still: you'll learn more about what μ can be by assessing what it's highly unlikely to be.

Table of Contents