SOC 3811/5811:
BASIC SOCIAL STATISTICS

Sampling Distributions

Sampling Distributions

Sampling Distribution

A theoretical distribution of means or proportions, taken from an infinite number of independent random samples of size n

Sampling distributions of means and proportions are normal regardless of the shape of the distribution of the variable that produces the mean or proportion

Central Limit Theorem

If n is sufficiently large, then the sample means from many random samples from a population with mean μ and variance σ^2 are approximately normally distributed with mean μ and variance

$$\sigma^2 \Big/ \sqrt{n}$$

Proportions

Proportions

Call p the population proportion

Call p-hat ($\hat{p}$) the sample proportion

If we generate many random sample of the same size, then the distribution of the several $\hat{p}$ will have a mean of p and variance of

$$\frac{\sigma^2}{\sqrt{n}} = \frac{\sum_{i=1}^{k}(Y_i - p)^2\, p_i}{\sqrt{n}} = \frac{(0-p)^2 p_0 + (1-p)^2 p_1}{\sqrt{n}} = \ldots = \frac{p(1-p)}{\sqrt{n}}$$

and standard deviation:

$$\sqrt{\frac{p(1-p)}{n}}$$

Proportions

But what if I select only **one** random sample, with one $\hat{p}$?

What is my best guess about p? It is $\hat{p}$

What is my best guess about the standard deviation of the sampling distribution of sample proportions, $\hat{p}$?

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

This is called the **standard *error*** of the sampling distribution of $\hat{p}$

Example

You randomly sampled 500 Minnesotans and found that 10% of them believe in UFOs. What is the probability that the population percentage of Minnesotans who believe in UFOs is between 9% and 11%?

Example

You randomly sampled 500 Minnesotans and found that 10% of them believe in UFOs. What is the probability that the population percentage of Minnesotans who believe in UFOs is between 9% and 11%?

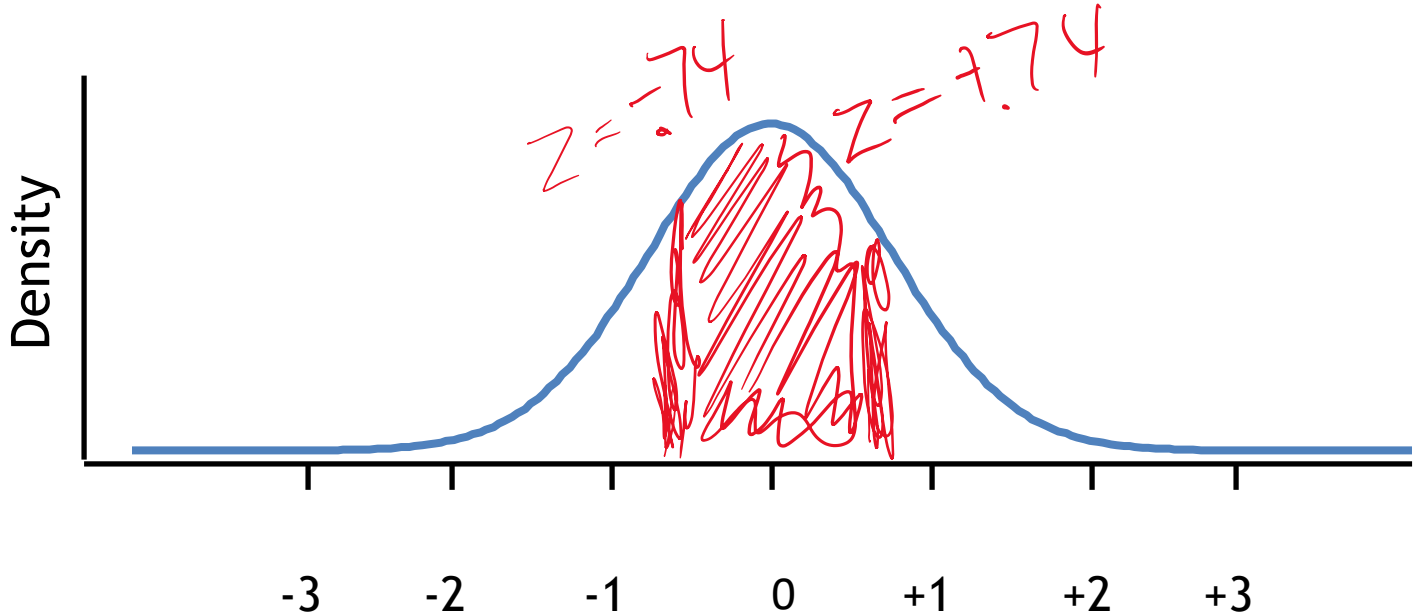
$$\text{Z for } 9\% = \frac{\hat{p}-p}{sd_p} = \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \frac{0.10-0.09}{\sqrt{\frac{0.10(0.90)}{500}}} = 0.74$$

$$\text{Z for } 11\% = \frac{\hat{p}-p}{sd_p} = \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \frac{0.10-0.11}{\sqrt{\frac{0.10(0.90)}{500}}} = -0.74$$

Example

You randomly sampled 500 Minnesotans and found that 10% of them believe in UFOs. What is the probability that the population percentage of Minnesotans who believe in UFOs is between 9% and 11%?

$P(-0.74 < Z < 0.74) = 0.7704 - 0.2296 = 0.5408$



WORKSHEET

You drew a random sample of 100 cats (because you wanted to have 100 cats on your farm) and observed that 15% of them are very mean cats. What is the probability that in the population of cats, between 10% and 20% are very mean?

WORKSHEET

You drew a random sample of 100 cats (because you wanted to have 100 cats on your farm) and observed that 15% of them are very mean cats. What is the probability that in the population of cats, between 10% and 20% are very mean?

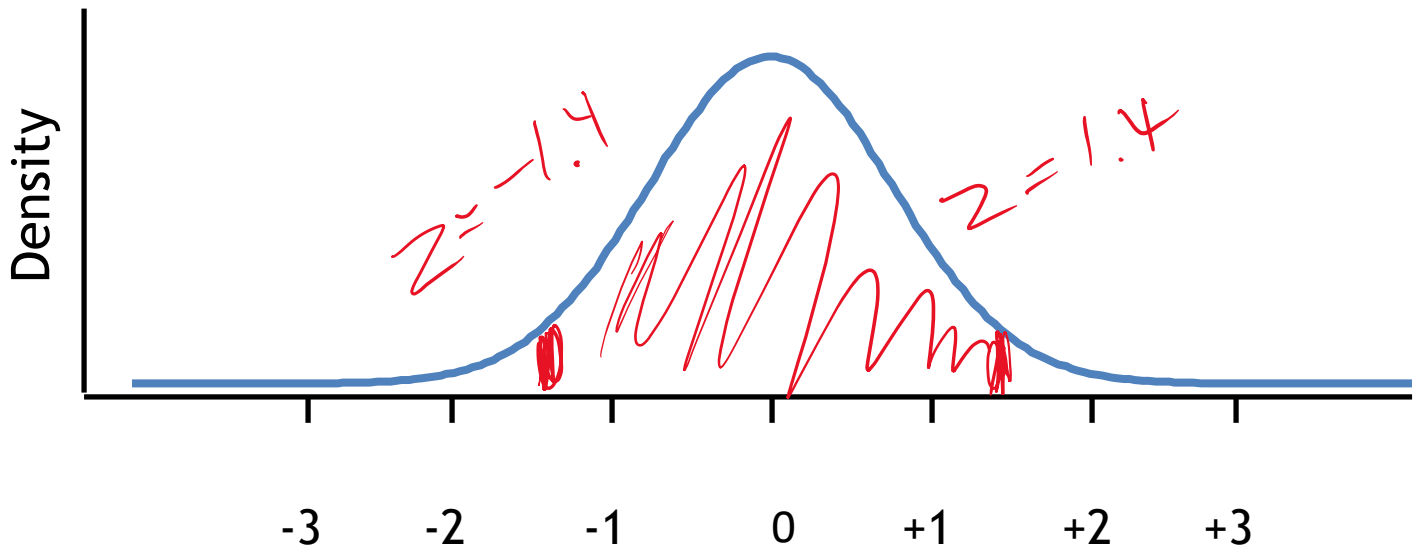
$$\text{Z for } 10\% = \frac{\hat{p}-p}{sd_p} = \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \frac{0.15-0.10}{\sqrt{\frac{0.15(0.85)}{100}}} = 1.4$$

$$\text{Z for } 20\% = \frac{\hat{p}-p}{sd_p} = \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \frac{0.15-0.20}{\sqrt{\frac{0.15(0.85)}{100}}} = -1.4$$

WORKSHEET

You drew a random sample of 100 cats (because you wanted to have 100 cats on your farm) and observed that 15% of them are very mean cats. What is the probability that in the population of cats, between 10% and 20% are very mean?

$P(-1.4 < Z < 1.4) = 0.9192 - 0.0808 = 0.8384$



Means

Means

The same sort of logic can be applied to the sampling distribution of sample means

In the population, the mean is μ and the standard deviation is σ

In any sample, the mean is $\bar{x}$ and the standard deviation is s

From the central limit theorem, the sampling distribution of means is centered over μ and has variance $\sigma^2 / \sqrt{n}$

Means

But what if I select only **one** random sample, with one $\bar{x}$?

What is my best guess about μ? It is $\bar{x}$

What is my best guess about the standard deviation of the sampling distribution of sample means, $\bar{x}$?

$$\sqrt{s^2/n}$$

This is called the **standard *error*** of the sampling distribution of $\bar{x}$

Example

You randomly selected 40 adult men from the population and observed their mean height to be 70 inches with a standard deviation of 5 inches. What is the probability that the mean height of men in the population is between 68 and 72 inches?

Example

You randomly selected 40 adult men from the population and observed their mean height to be 70 inches with a standard deviation of 5 inches. What is the probability that the mean height of men in the population is between 68 and 72 inches?

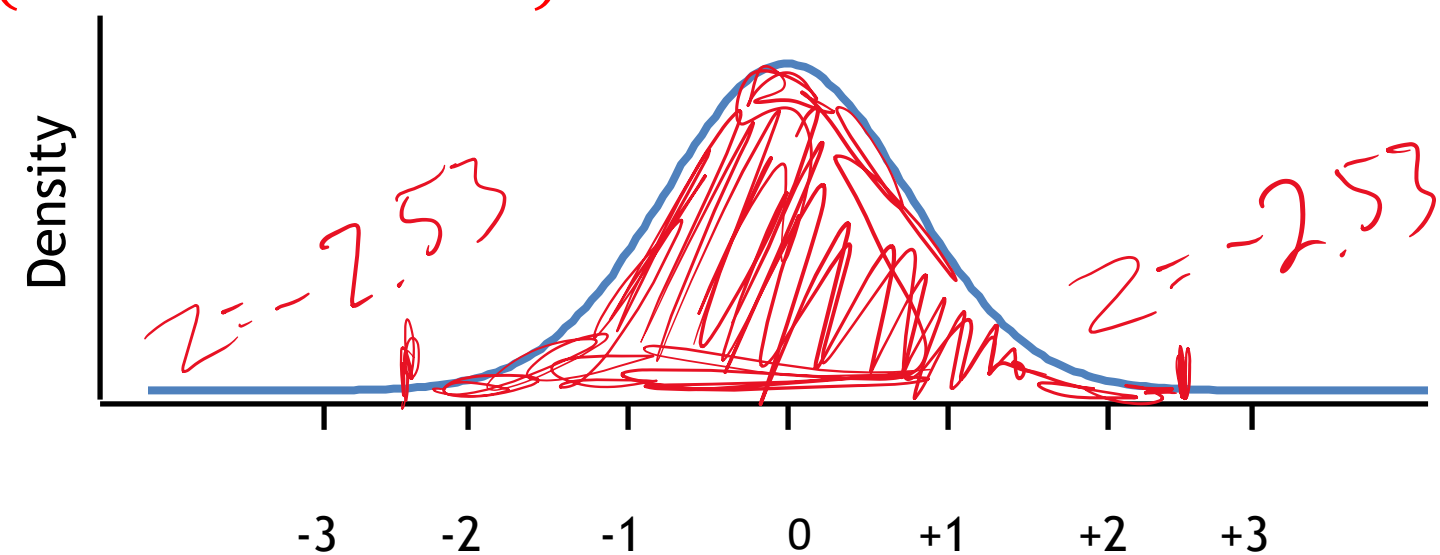
$$\text{Z for 68"} = \frac{\bar{x}-\mu}{sd} = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}}} = \frac{70-68}{\sqrt{\frac{5^2}{40}}} = 2.53$$

$$\text{Z for 72"} = \frac{\bar{x}-\mu}{sd} = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}}} = \frac{70-72}{\sqrt{\frac{5^2}{40}}} = -2.53$$

Example

You randomly selected 40 adult men from the population and observed their mean height to be 70 inches with a standard deviation of 5 inches. What is the probability that the mean height of men in the population is between 68 and 72 inches?

$P(-2.53 < Z < 2.53) = 0.9943 - 0.0057 = 0.9886$



t distribution

When n is large … say, more than 50 … sampling distributions of means follow the Z distribution

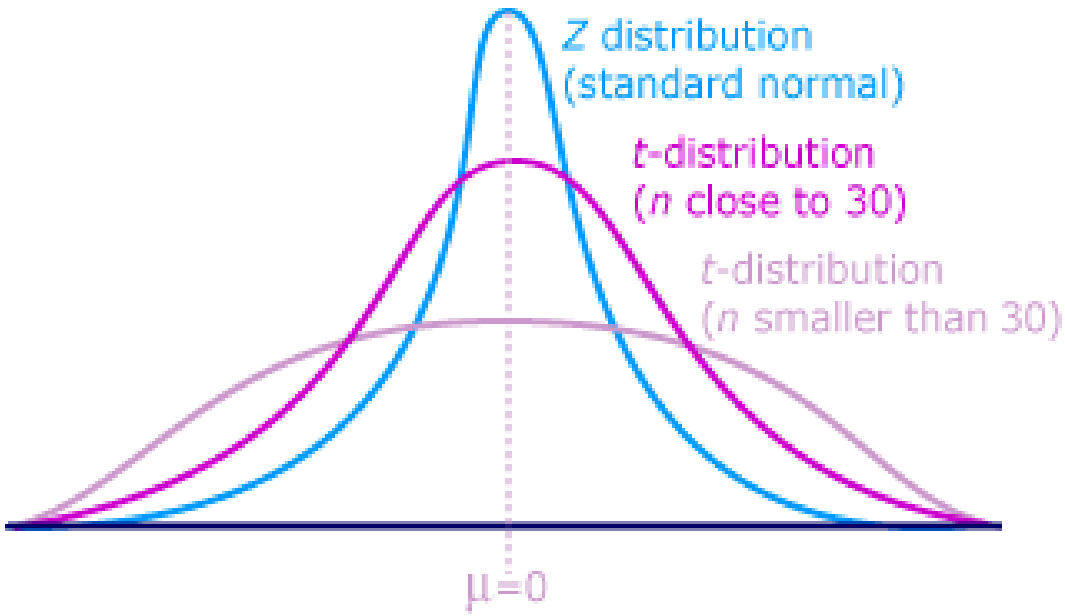
When n is smaller … say, less than 50 … sampling distributions of means follow **t distributions**, not Z distributions

There is a different t distribution for each value of n; each t distribution is defined by its degrees of freedom, df, where df equal n-1

t distribution

t scores are computed the same way as Z scores

$$Z = \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}} \qquad t = \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}}$$



t distribution

When n is large, the t and the Z distributions are (approximately) the same and so the area under the curve within any given range of Z or t scores is the same

When n is small, use the t distribution with n-1 degrees of freedom

To be safe, in practice most people always use the t distribution for sampling distributions of means

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

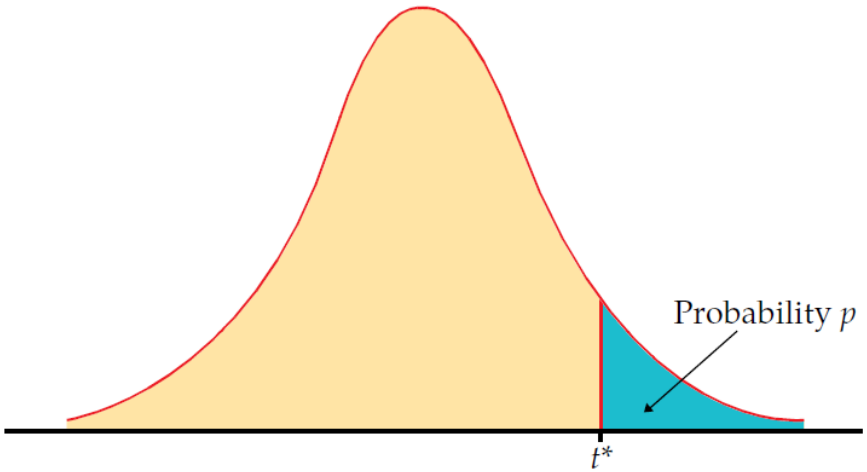


TABLE D

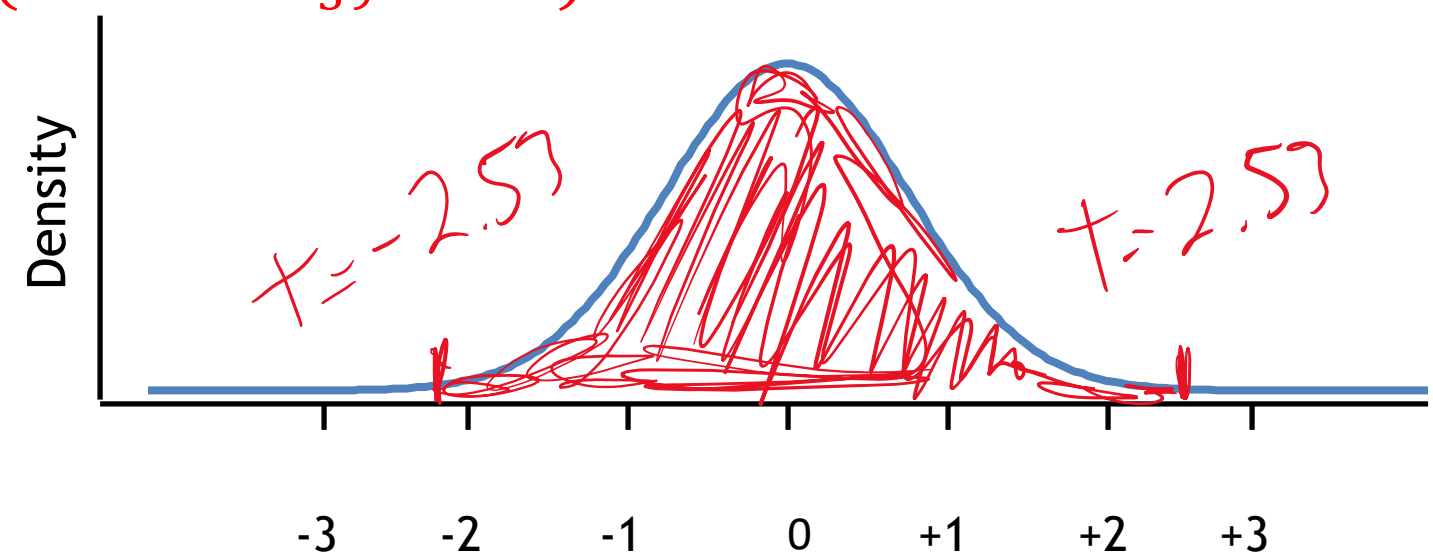
t distribution critical values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965

Example

You randomly selected 40 adult men from the population and observed their mean height to be 70 inches with a standard deviation of 5 inches. What is the probability that the mean height of men in the population is between 68 and 72 inches?

$P(-2.53 < t_{39} < 2.53) =$ about 0.99



WORKSHEET

You randomly selected 35 people and observed that they go to the bathroom an average of 4.5 times per day with a standard deviation of 1.0 times. What is the probability that the population mean number of times that people go to the bathroom per day is between 3.9 and 5.1?

WORKSHEET

You randomly selected 35 people and observed that they go to the bathroom an average of 4.5 times per day with a standard deviation of 1.0 times. What is the probability that the population mean number of times that people go to the bathroom per day is between 3.9 and 5.1?

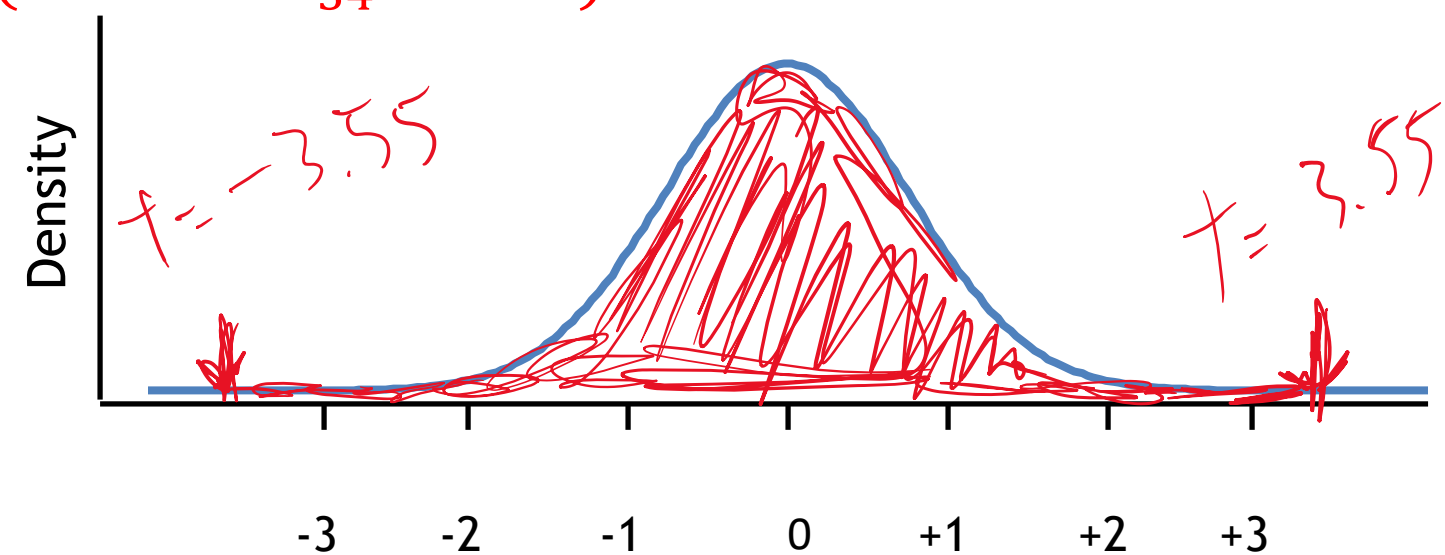
$$\text{t for } 3.9 = \frac{\bar{x}-\mu}{sd} = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}}} = \frac{4.5-3.9}{\sqrt{\frac{1^2}{35}}} = 3.55$$

$$\text{t for } 5.1 = \frac{\bar{x}-\mu}{sd} = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}}} = \frac{4.5-5.1}{\sqrt{\frac{1^2}{35}}} \text{-}3.55$$

WORKSHEET

You randomly selected 35 people and observed that they go to the bathroom an average of 4.5 times per day with a standard deviation of 1.0 times. What is the probability that the population mean number of times that people go to the bathroom per day is between 3.9 and 5.1?

$P(-3.55 < t_{34} < 3.55) =$ between 0.01 and 0.02



Why is this Useful?

1. Confidence Intervals

Based on the distribution of Y in sample data, we are confident that the distribution of Y in the population has particular qualities
(e.g., that its mean is within a certain range of values)

"With 95% certainty, I conclude based on my sample data that between 25% and 35% of everyone in the population has been arrested"

Why is this Useful?

2. Hypothesis Tests

Based on the distribution of Y in the sample data, we can evaluate the likely truth of theoretically-informed hypotheses about the distribution of Y in the population (e.g., that the mean of X is above some value)

"With 95% certainty, I reject the claim that fewer than 20% of everyone in the population has ever been arrested"