Today: Finish Chapter 9 (Sections 9.6 to 9.8 and 9.9 Lesson 3)

ANNOUNCEMENTS:
- Quiz #7 begins after class today, ends Friday at noon.
- Jason Kramer will give the lecture on Friday.
- Please check your grades on eee and let me know if there are problems.

HOMEWORK: (Due Friday)
Chapter 9: #66, 90, 91

Update on the five situations we will cover for the rest of this quarter:

Parameter name and description	**Population parameter**	**Sample statistic**
***For Categorical Variables:* [Done!]**		
One population proportion (or probability)	p	$\hat{p}$
Difference in two population proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$
***For Quantitative Variables:* [Today, Fr, M]**		
One population mean	μ	$\overline{x}$
Population mean of paired differences (dependent samples, paired)	μ_d	$\overline{d}$
Difference in two population means (independent samples)	$\mu_1 - \mu_2$	$\overline{x}_1 - \overline{x}_2$

For each situation will we:

- Learn about the *sampling distribution* for the sample statistic
- Learn how to find a *confidence interval* for the true value of the parameter
- *Test hypotheses* about the true value of the parameter

Recall, general format for all sampling distributions in Ch. 9:

The sampling distribution of the sample statistic is approximately normal, with:

- Mean = population parameter (p, $p_1 - p_2$, μ, etc.)
- Standard deviation = *standard deviation of* _______; the blank is filled in with the statistic ($\hat{p}$, $\hat{p}_1 - \hat{p}_2$, $\bar{x}$ etc.)
- Often the standard deviation must be estimated, and then it is called the *standard error of* ________.

See summary table on pages 382-383 for all details!

Today: Sampling distributions for:

- one mean
- mean difference for paired data
- difference between means for independent samples

Remember, two samples are called **independent samples** when the measurements in one sample are not related to the measurements in the other sample. Could come from:

- Separate samples
- One sample, divided into two groups by a categorical variable (such as male or female)
- Randomization into two groups where each unit goes into only one group

Paired data occur when two measurements are taken on the same individuals, or individuals are paired in some way.

Sampling Distribution for a Sample Mean (Section 9.6)

Suppose we take a random sample of size n from a population and measure a quantitative variable.

Notation:

μ = mean for the *population* of measurements.

σ = standard deviation for the *population* of measurements.

$\bar{x}$ = *sample* mean for a random sample of n individuals.

s = *sample* standard deviation for the random sample

The sampling distribution of the sample *mean* $\overline{x}$ is approximately normal, with:

- Mean = population parameter = μ
- Standard deviation = *standard deviation of* $\overline{x}$ =

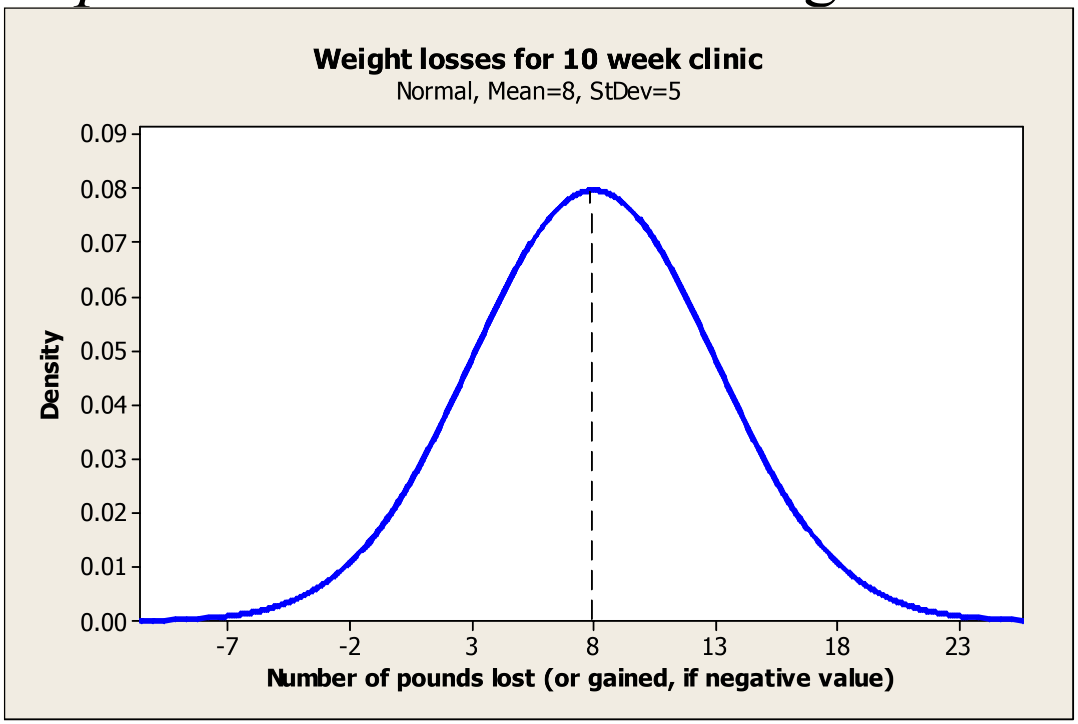
$$s.d.(\overline{x}) = \frac{\sigma}{\sqrt{n}}$$

- Often the standard deviation must be estimated, and then it is called the *standard error of* $\overline{x}$. Replace σ with the sample standard deviation, so

$$s.e.(\overline{x}) = \frac{s}{\sqrt{n}}$$

Suppose we want to estimate the ***mean*** *weight loss* for the *population* of people who attend weight loss clinics for 10 weeks. *Suppose* the distribution of weight losses is approximately normal, $\mu = 8$ pounds, $\sigma = 5$ pounds. (Empirical rule: see picture)

Population of *individual* weight losses



- We plan to take a random sample of 25 people from this population and record weight loss for each person, then find sample mean $\bar{x}$.

- We know the value of the sample mean will vary for different samples of n = 25. How much will they vary? Where is the center of the distribution of possibilities?

Results for four possible random samples of 25 people, with the corresponding sample mean $\bar{x}$ and sample standard deviation s:

Sample 1: $\bar{x}$ = 8.32 pounds, s = 4.74 pounds.
Sample 2: $\bar{x}$ = 6.76 pounds, s = 4.73 pounds.
Sample 3: $\bar{x}$ = 8.48 pounds, s = 5.27 pounds.
Sample 4: $\bar{x}$ = 7.16 pounds, s = 5.93 pounds.

Note:

- Each sample had a different sample mean, which did not always match the population mean of 8 pounds.
- Although we cannot determine whether one sample mean will accurately reflect the population mean, statisticians have determined what to expect for *all possible* sample means.

μ = mean for population of interest = 8 pounds
σ = standard deviation for population of interest = 5 pounds.
$\bar{x}$ = sample mean for a random sample of *n* individuals.

Then the *sampling distribution of* $\bar{x}$ is approximately normal, with

- $Mean = \mu$
- $Standard\ deviation = s.d.(\bar{x}) = \frac{\sigma}{\sqrt{n}}$

Example: Mean of 25 weight losses, the distribution of possible values is approximately normal with:

- mean = 8 pounds
- standard deviation = $\frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{25}} = 1$ pound

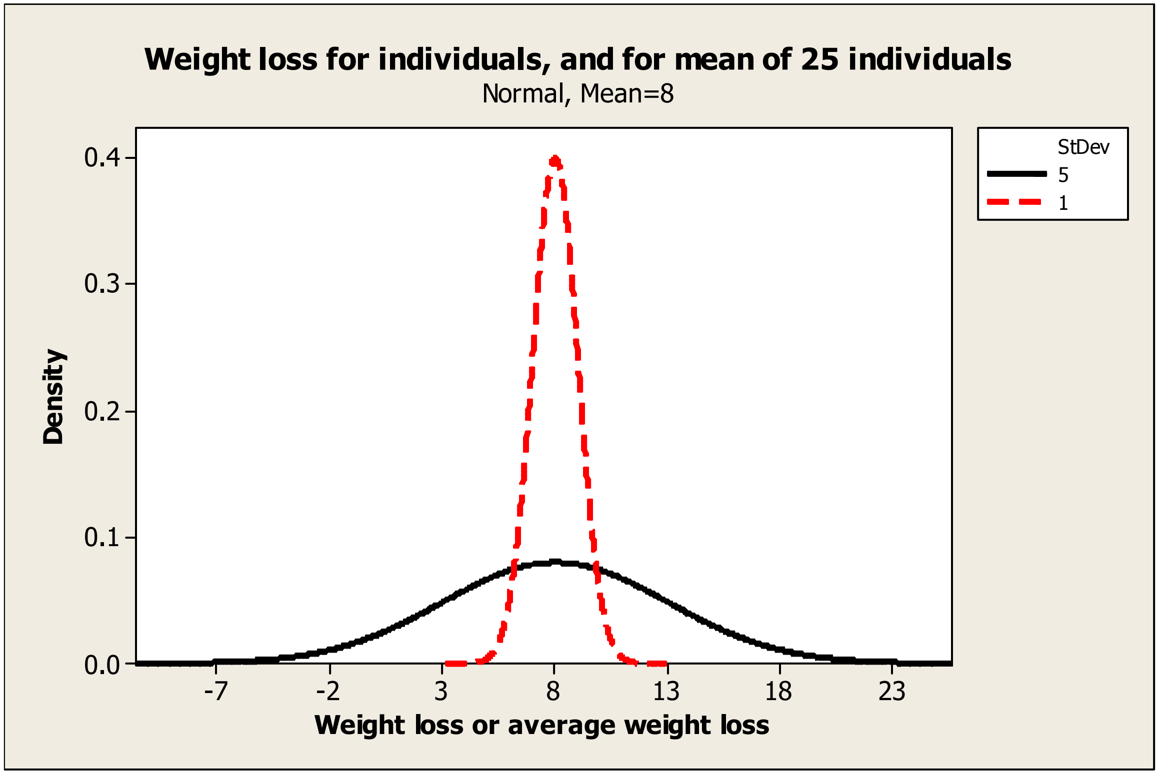
Compare: *individual* weight loss, $\bar{x}$ for $n = 25$, $\bar{x}$ for $n = 100$

	Individual weight loss	Mean of 25	Mean of 100
Mean	8 pounds	8 pounds	8 pounds
St. Dev.	5 pounds	1 pound	½ pound

Conditions for sampling distribution of $\bar{x}$ to be approximately normal:

- *Population* (individual values) are approx. bell-shaped OR
- Sample size is *large* (at least 30, more if outliers)

Comparing original population with sampling distribution of $\bar{x}$:

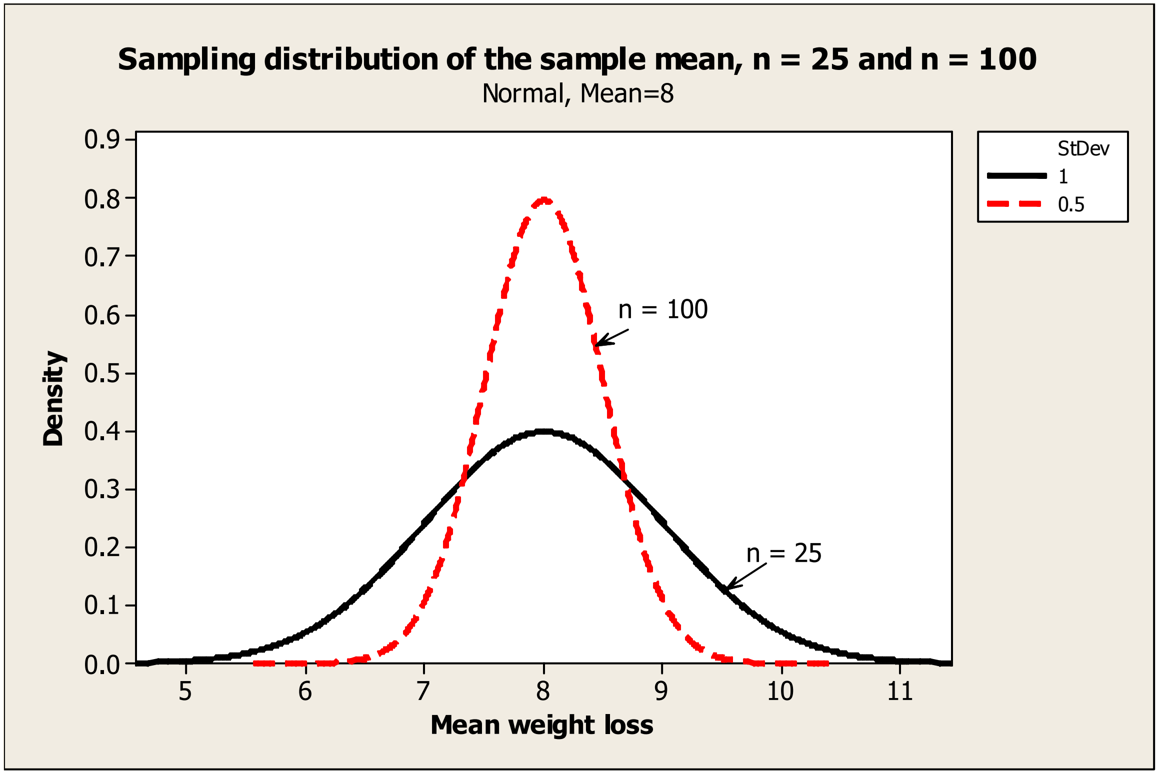


From the empirical rule:

	68%	95%	99.7%
Individuals	3 to 13 pounds	-2 to 18 pounds	-7 to 23 pounds
Mean of $n = 25$	7 to 9 pounds	6 to 10 pounds	5 to 11 pounds

Note that larger sample size will result in *smaller* s.d.($\bar{x}$)

Compare sampling distribution for n = 25 and n = 100:



In other words, for *larger* samples, $\bar{x}$ will be closer to μ in general, and thus will be a better estimate for μ.

Example where the original population is *not* bell-shaped:

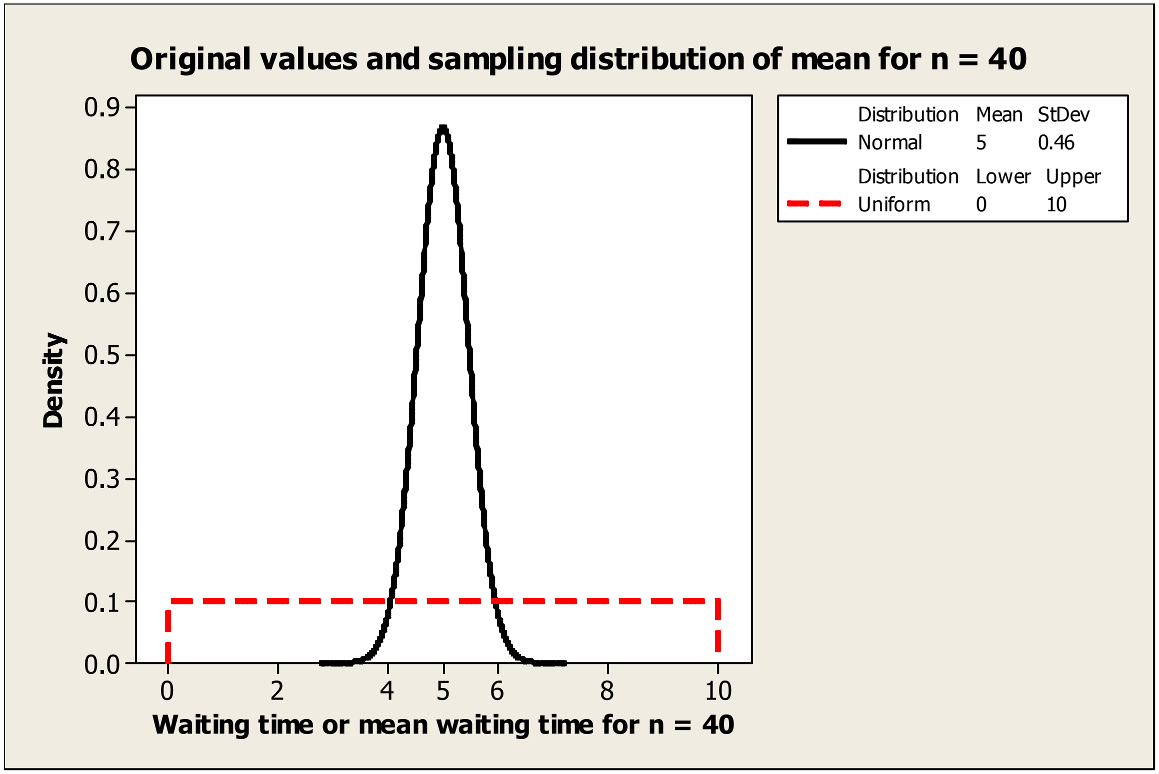
A bus runs every 10 minutes. When you show up at the bus stop, it could come immediately, or anytime up to 10 minutes. So the time you wait for it is *uniform*, from 0 to 10 minutes, and independent from day to day.

Population mean = μ = 5 minutes, population s.d. = σ = $\sqrt{\frac{10^2}{12}}$ = 2.9

What is the sampling distribution for $\bar{x}$ for n = 40 days?

Even though the *original times* are *uniform* (flat shape), the possible values of the sample mean $\bar{x}$ are:

- Approximately normal
- Mean = 5 minutes
- Standard deviation = $\frac{\sigma}{\sqrt{n}} = \frac{2.9}{\sqrt{40}}$ = 0.46 minutes



Examples of possible samples:

6.9	6.2	8.8	8.3	7.1	6.5	7.3	9.5	3.4	9.9	5.8	1.4	3	4.1	4.4	0.6	$\bar{x} = 5.29$
2.7	1.2	7.4	0.7	6.8	7.7	6.2	6.1	3.3	1	5.3	9.4	1	0.8	1	9.4	
8.1	3.9	7.2	8.6	1.1	0.4	9.9	9.2									
0.6	3.2	0.8	0.2	8.5	1.4	4.7	0.5	9.7	8.9	6.3	3.3	0.8	4.1	2.6	3.7	$\bar{x} = 4.3$
5.7	3.2	8.9	2.3	1.1	9.9	3	0.8	7.9	0.8	5.9	2.5	7.9	7.6	4	2.2	
0.6	0.1	6.1	6.9	8.1	2.6	9.6	5.3									

Sections 9.7 and 9.8: Sampling distributions for *mean of paired differences*, and for *differences in means for independent samples*
Need to learn to distinguish between these two situations.

Notation for paired differences:

- d_i = difference in the two measurements for individual $i = 1, 2, ..., n$
- μ_d = mean for the *population* of differences, if all possible pairs were to be measured
- σ_d = the standard deviation for the *population* of differences
- $\bar{d}$ = the mean for the *sample* of differences
- s_d = the standard deviation for the *sample* of differences

Example: IQ measured after listening to Mozart and to silence
d_i = difference in IQ for student i for the two conditions
μ_d = *population* mean difference, if all students measured (unknown)
$\bar{d}$ = the mean for the *sample* of differences = 9 IQ points
Based on *sample*, we want to *estimate* mean *population* difference

Notation for difference in means for independent samples:

μ_1 = *population* mean for the first population
μ_2 = *population* mean for the second population
Parameter of interest is $\mu_1 - \mu_2$ = the difference in *population* means

$\bar{x}_1$ = *sample* mean for the sample from the first population
$\bar{x}_2$ = *sample* mean for the sample from the second population
The sample statistic is $\bar{x}_1 - \bar{x}_2$ = the difference in *sample* means

σ_1 = *population* standard deviation for the first population
σ_2 = *population* standard deviation for the second population

s_1 = *sample* standard deviation for the sample from the 1^{st} population
s_2 = *sample* standard deviation for the sample from the 2^{nd} population

n_1 = size of the sample from the 1^{st} population
n_2 = size of the sample from the 2^{nd} population

Examples where independent samples might be used:

- Compare weight loss for men and women at the weight loss clinic.
- Compare UCI students with students from another campus on quantitative measures like hours spent studying per week, income, etc
- Compare number of sick days off from work for people who had a flu shot and people who didn't
- Compare change in blood pressure for people randomly assigned to a meditation program or an exercise program for 3 months.

Examples where paired data might be used:

- Estimate average difference in income for husbands and wives
- Compare SAT scores before and after a training program
- Weight loss example can be thought of as paired difference, with weight before and weight after the program

Note that paired differences are similar to the "one mean" situation, except special notation tells us that the means are for differences.

Conditions for the sampling distributions for these two situations are the same as for a single mean, with a slight twist:

- For paired differences, population of *differences* must be bell-shaped OR sample must be large.
- For difference in means for independent samples, *both* populations must be bell-shaped OR *both* sample sizes must be large.

In both cases, the sampling distribution for the sample statistic is approximately normal, with mean = population parameter of interest.

For paired differences: s.d.($\bar{d}$) = $\frac{\sigma_d}{\sqrt{n}}$ (same as one mean, but with d's)

For difference in two means: s.d.($\bar{x}_1 - \bar{x}_2$) = $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Standardized Statistics:

For all 5 cases in Chapter 9, as long as the conditions are satisfied for the sampling distribution to be approximately normal, the **standardized statistic** for a sample statistic is:

$$z = \frac{\text{sample statistic - population parameter}}{\text{s.d.(sample statistic)}}$$

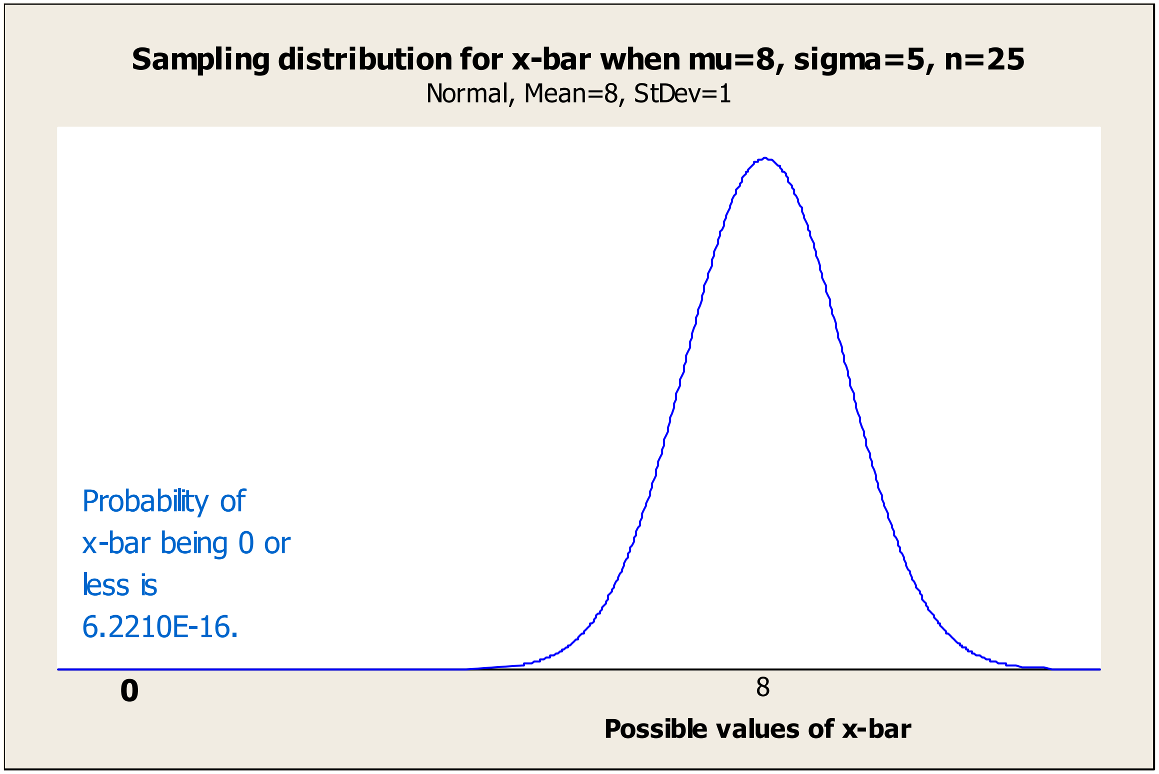
Note that the denominator has s.d., *not* s.e.

For one mean:

$$z = \frac{\bar{x} - \mu}{s.d.(\bar{x})} \quad = \quad \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad = \quad \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}$$

Example of weight loss clinic. Suppose in a sample of 25 clients the average weight loss is 0.

If *population* mean weight loss is really 8 pounds with σ = 5 pounds, how unlikely is a *sample* mean of 0 pounds for n = 25?



How to compute this answer:

Sample means for $n = 25$ are approximately normal with mean of 8 pounds and s.d. of 1 pound. So, the standardized score for 0 is:

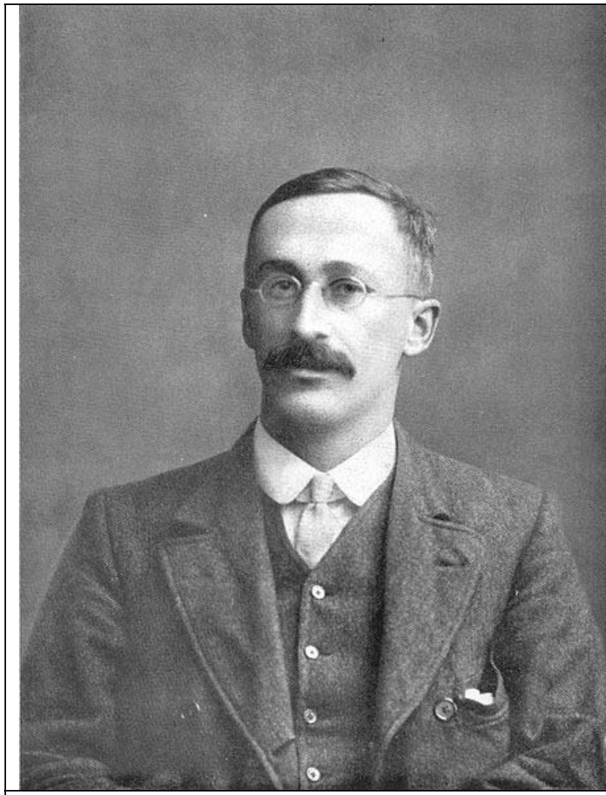
$$z = \frac{\sqrt{25}(0-8)}{5} = -8$$

A z-score of –8 is not very likely! So if we saw a 0 mean weight loss, we would not believe that the population mean is 8 pounds!

Note: When σ is not known, we must use the sample standard deviation s instead. <u>Standard error</u> of $\bar{x}$ is $s.e.(\bar{x}) = \frac{s}{\sqrt{n}}$

In that case, *standardized statistic* has a *t-distribution*, also called *Student's t distribution*.

Student's *t* distribution



In 1908 William Sealy Gossett figured out the formula for the *t* distribution. Called Student's *t* because… explained in class!

Standardized Statistic Using Standard Error

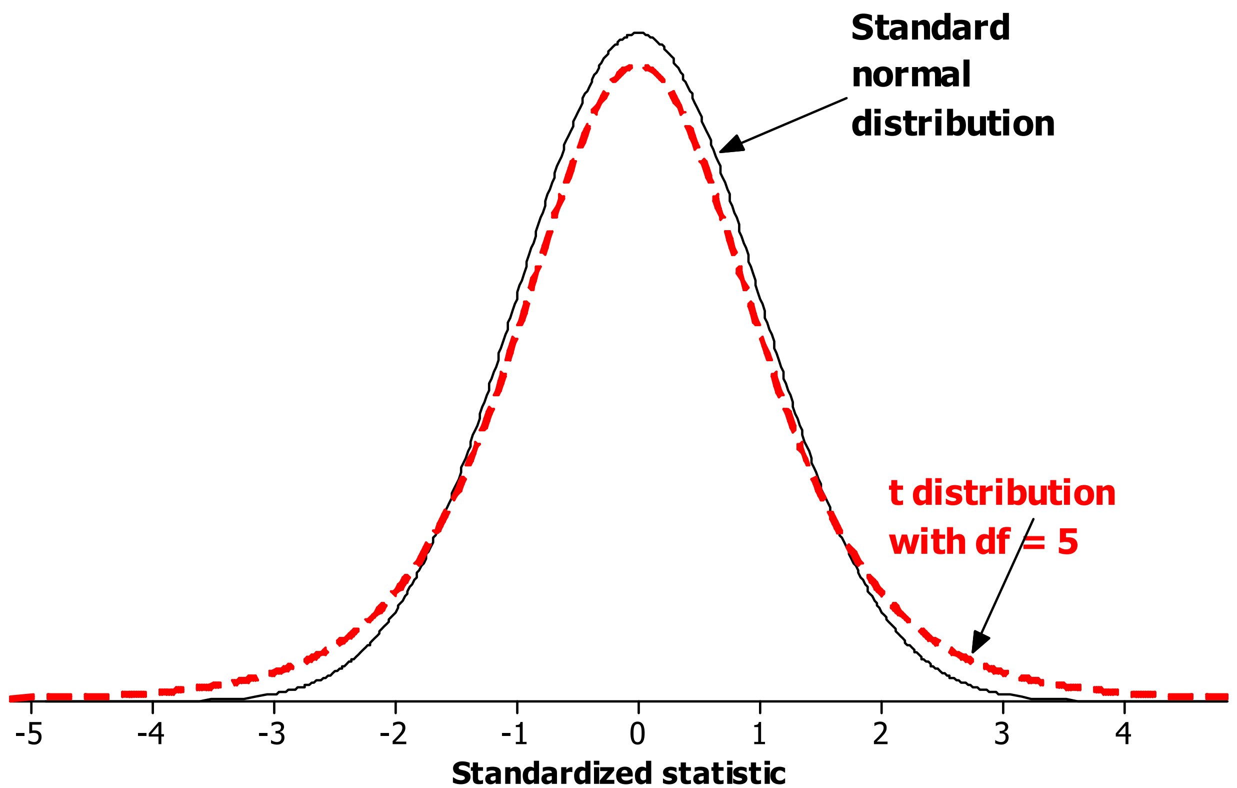
Usually we don't know σ (population standard deviation), so we need to use *s* (sample standard deviation). In that case, the standardized statistic for $\bar{x}$ is

$$t = \frac{\bar{x} - \mu}{s.e.(\bar{x})} = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{\sqrt{n}(\bar{x} - \mu)}{s}$$

This has a *Student's t distribution* with degrees of freedom = *n* – 1

- It looks almost exactly like the normal distribution
- It is completely specified by knowing the "df"
- It gets closer and closer to the normal distribution, and when degrees of freedom = infinity, it is exactly the normal distribution.

Comparison of *t* distribution with df = 5 and standard normal distribution



For example, middle 95% for *t* with df = 5 is −2.57 to +2.57
For standard normal, it is about −2 to + 2
In Chapter 11 (Friday) we will learn how to find probabilities.

Summary of sampling distributions for the 5 parameters (p. 382):

- The *statistic* has a sampling distribution.
- It is *approximately normal* if the sample(s) is (are) large enough.
- The *mean of the sampling distribution* = the parameter.
- The *standard deviation of the sampling distribution* is in the table below, in the column "standard deviation of the statistic."
- Sometimes it needs to be estimated, then "*standard error*" is used.

	Parameter	Statistic	Standard Deviation of the Statistic	Standard Error of the Statistic	Standardized Statistic with s.e.
One proportion	p	$\hat{p}$	$\sqrt{\frac{p(1-p)}{n}}$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	z
Difference Between Proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	z
One Mean	μ	$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$	t
Mean Difference, Paired Data	μ_d	$\bar{d}$	$\frac{\sigma_d}{\sqrt{n}}$	$\frac{s_d}{\sqrt{n}}$	t
Difference Between Means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	t

	Para-meter	Sta-tistic	Standard Deviation of the Statistic	Standard Error of the Statistic	z or t? (with s.e.)
One proportion	p	$\hat{p}$	$\sqrt{\frac{p(1-p)}{n}}$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	z
Difference Between Proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	z
One Mean	μ	$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$	t
Mean Difference, Paired Data	μ_d	$\bar{d}$	$\frac{\sigma_d}{\sqrt{n}}$	$\frac{s_d}{\sqrt{n}}$	t
Difference Between Means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	t