ECON 3818

Chapter 20

Kyle Butts

24 August 2021

Chapter 20: Inference about a Population Mean

Conditions for Inference about a Mean

We've discussed using confidence intervals and tests of significance for the mean μ of a population

In general, our analysis relied on two conditions:

- The data is from a **simple random sample** from the population
- Observations from the population have a **normal distribution** with a mean, μ, which is generally unknown and a variance σ^2, which we've been assuming is known.

We'll now talk about the situation where μ and σ^2 are both unknown.

Important Reminder

For a sample mean $\bar{X}_n$, the sampling distributino has the following variance and standard deviation:

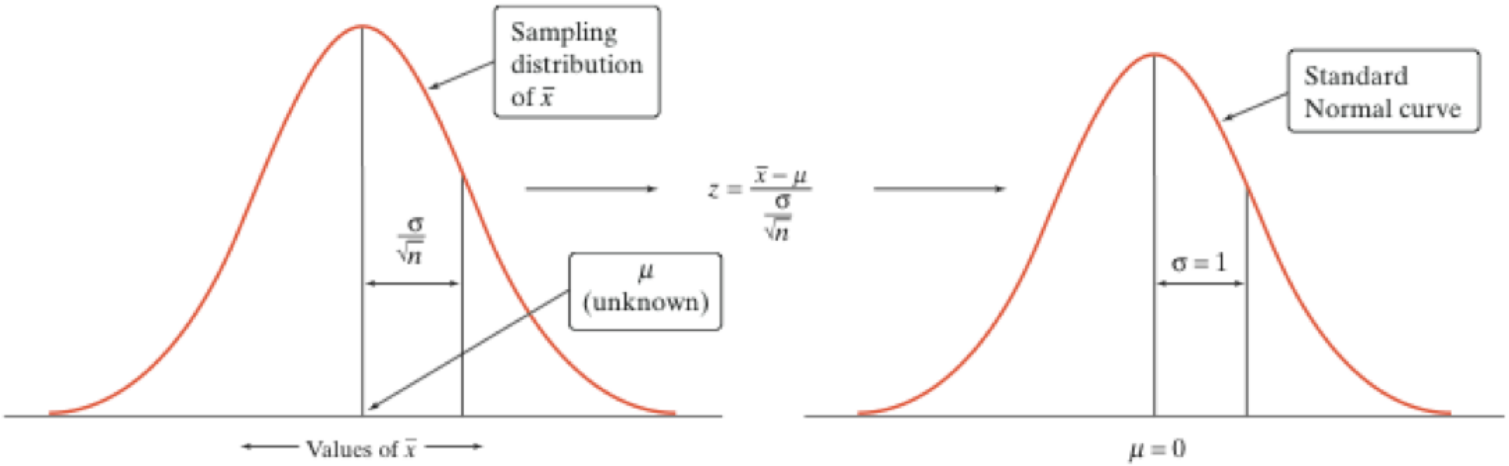
$$\text{Variance} = \frac{\sigma^2}{n}$$

$$\text{Standard Deviation} = \frac{\sigma}{\sqrt{n}}$$

When σ^2 is Known

Whenever we know the population variance, σ^2, we base confidence intervals and tests for μ with the z-statistic:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \implies Z \sim N(0,1)$$



When σ^2 is Unknown

When we don't know the true variance, we must estimate it using the sample variance s^2. Again, since we're talking about sample means, the standard deviation of a sample mean, based of a *population with unknown variance* is:

$$\frac{\sigma^2}{n} \text{ which is estimated by } \frac{s^2}{n}$$

$\frac{s}{\sqrt{n}}$ is called the **standard error**.

When σ^2 is Unknown

When we don't know σ, we estimate it by the *sample standard deviation*, s. What happens when we standardize?

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim \quad ?$$

This new statistic does not have a normal distribution!

The t Distribution

When we standardize based on the sample standard deviation, s, our statistic has a new distribution called a t-distribution

Consider a simple random sample of size n, the t-statistic:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has the t-distribution with $n - 1$ degrees of freedom

The t-statistic has the same interpretation as z-statistic: it says how far away $\bar{X}$ is from its mean μ, in standard deviation units.

- The distribution is different because we have one aditional source of *uncertainty*: we estimate the standard deviation with s.

The t-Distribution

There is a different t distribution for each sample size, specified by its degrees of freedom

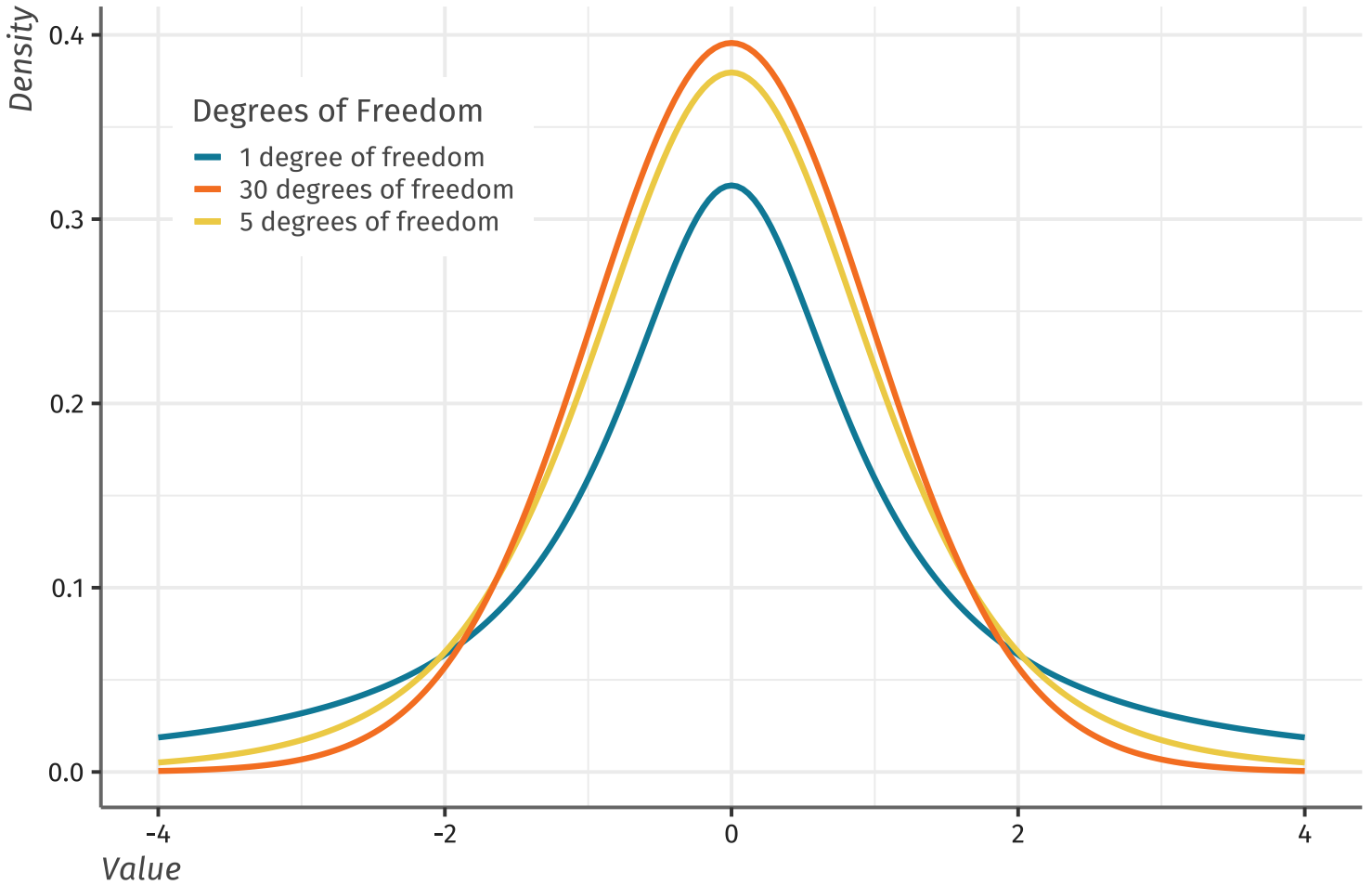
- We denote the t distribution with $n-1$ degrees of freedom as t_{n-1}
- The curve has a different shape than the standard normal curve
    - It is still symmetric about it's peak at 0
    - But, it has more area in the tails.

"More Area in the Tails"?

Example: Let's say I want to predict the height of the class by selecting $n = 3$ people and taking their mean. Sometimes I will pick 3 basketball players or 3 short people.

- If I am using $n = 20$ people, then it is much more rare to select only tall or only short people (this means that extreme sample means are more rare).
- I might still select the tallest or shortest person in the class, but they're only one of 20 instead of one of 3.

Graphs of t-Distributions



Graphs of t-Distributions

Greater degree of dispersion in the t distribution

- It has an additional source of random variability in the sample standard deviation s

The variance from estimating s decreases when the sample size increases

- This means the t-distribution looks more like the z-distribution as sample size grows to infinity
- Once degrees of freedom exceeds 31, the t-distribution is close enough to standard normal to approximate probabilities using the z-distribution

Finding t-values

Finding t-values

As you see, once we have more than 30 degrees of freedom, we can use the z-scores

23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905

Confidence Interval with t-Distribution

Calculating a confidence interval:

If the variance, σ^2, is known:

- Use Z-distribution

If the variance, σ^2, is unknown:

- 31 or more observations in sample, use Z-distribution
- 30 or fewer observations in sample, use t-distribution

Example

Say you want to construct a 95% confidence interval using 20 observations from population with unknown μ and σ, with a sample mean, $\bar{X} = 22.21$ and a sample standard deviation, $s = 1.963$.

σ is unknown and $n \leq 31 \implies$ use t-distribution:

$$CI = \bar{X} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the critical value:

$$ t^{\frac{1-C}{2}}_{df} = t_{19}^{0.025} $$

Example

Find $t_{19}^{0.025}$ using t-distribution table:

df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534

$$t_{19}^{0.025} = 2.093$$

Example

$$CI = \bar{X} \pm t^* \frac{s}{\sqrt{n}}$$

$$CI = 22.21 \pm 2.093 \cdot \frac{1.963}{\sqrt{20}}$$

Which means the confidence interval is [21.29, 23.13].

We are 95\% confident that the population mean is between 21.29 and 23.13.

Clicker Question

Which table would you use for the following confidence interval?

99% confidence interval with $n = 1000$ observations

a. z-table

b. t-table

Clicker Question

What critical value t^* would you use for the following confidence interval?

90% confidence interval with $n = 2$ observations

a. 2.92

b. 6.31

c. 3.08

d. 1.89

df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534

Clicker Question

What critical value t^* would you use for the following confidence interval?

95% confidence interval with $n = 16$ observations

a. 1.753

b. 1.745

c. 2.13

d. 2.12

df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534

Clicker Question -- Midterm Example

A randomly sampled group of patients at a major U.S. regional hospital became part of a nutrition study on dietary habits. Part of the study consisted of a 50-question survey asking about types of foods consumed. Each question was scored on a scale from one (most unhealthy behavior) to five (most healthy behavior). The answers were summed and averaged. The population of interest is the patients at the regional hospital. A sample of $n = 15$ patients produced the following statistics $\bar{X} = 3.3$ and $s = 1.2$. A 99% confidence interval is given by:

a. $(2.37, 4.22)$.
b. $(2.64, 3.97)$.
c. $(2.69, 3.91)$.
d. $(2.5, 4.1)$.

Hypothesis Test with t-Distribution

Like the confidence interval, the t test is very similar to the Z test introduced earlier.

If we have SRS of size n from population with unknown μ and σ. To test the hypothesis: $H_0 : \mu = \mu_0$, compute the t-statistic:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

We then use this t-statistic to calculate the p-value

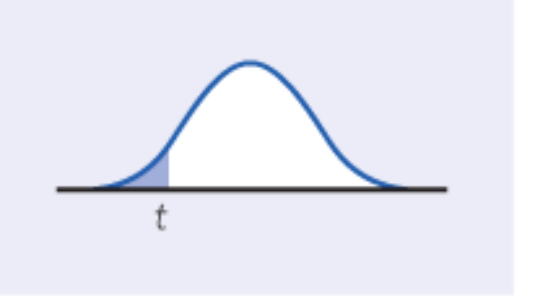
- The p-values are exact if the population does happen to be normally distributed
- Otherwise, they are approximately correct for large enough n

p-Values in t-Distribution

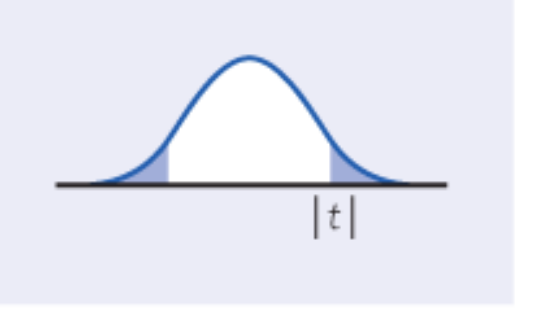
In terms of a variable T having the t_{n-1} distribution, the P-value for a test of H_0 against

$H_a: \mu > \mu_0$ is $P(T \geq t)$

$H_a: \mu < \mu_0$ is $P(T \leq t)$



$H_a: \mu \neq \mu_0$ is $2P(T \geq |t|)$



Example

Suppose we have a sample of 12 observations and calculate a sample mean $\bar{X} = 113.75$ and a sample standard deviation, $s = 93.90$, and we want to test the following hypothesis at the $\alpha = 0.10$ level:

$$H_0 : \mu = 88$$

$$H_1 : \mu > 88$$

The set up is similar to how we've done hypothesis testing before, but we must use the t distribution.

$$P(\bar{X} > 113.75 \mid \mu = 88)$$

Example

When we "standardize", we are going to be using the t-distribution since we don't know the population variance and we have a small sample.

$$\begin{aligned} P(\bar{X} > 113.75 \mid \mu = 88) &= P(\frac{\bar{X} - \mu_0}{s/\sqrt{n}} > \frac{113.75 - 88}{93.90/\sqrt{12}}) \\ &= P(t_{11} > 0.95) \end{aligned}$$

Example

Our next step is to find the p-value associated with that t-statistic

We can pin down the p-value between two points by using the t-table used earlier.

- Look at the row, df=11
- See which columns 0.95 lies between, and look at their associated probabilities

Example

df/p	0.40	0.25	0.10	0.05	0.025	0.01
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100

We see that along the row with 11 degrees of freedom, Our t-statistic (0.95) is between 0.25, which has a t-statistic of 0.697, and 0.10, which has a t-statistic of 1.36.

We conclude that $0.10 \leq p - \text{value} \leq 0.25$

Example

Our level of significance is $\alpha = 0.10$, and we concluded that the p-value associated with this hypothesis is test is greater than 0.10.

Therefore:

$$\alpha \leq p - \text{value} \implies \text{Do not reject } H_0$$

That main difference between solving a hypothesis test with z-statistic versus t-statistic is that we will generally only be able to find a range for the p-value with t-statistic, whereas we found an exact value with a z-statistic.

- You would need 30 seperate t tables to have matching exactness of Z-table.

Clicker Question

Say we have calculated a t-statistic of 1.6 from a sample of 24 observations. Do we reject at the $\alpha = 0.10$ level? What about the $\alpha = 0.05$ level?

a. Reject at both $\alpha = 0.05$ and $\alpha = 0.10$
b. Do not reject at both $\alpha = 0.05$ and $\alpha = 0.10$
c. Reject at $\alpha = 0.05$, but do not reject at $\alpha = 0.10$
d. Reject at $\alpha = 0.10$, but do not reject at $\alpha = 0.05$

Hypothesis Testing with t-Distribution

When we conduct a hypothesis test we follow these steps:

1. Calculate the test statistic

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

2. Determine range of p-values associated with test statistic using t-table
    - Row is determined by degrees of freedom, $n - 1$
    - Find which two columns your t-statistic lies between
3. Similar to hypothesis testing with a Z distribution, if using a two-tailed test, multiply p-value by 2.
4. Compare range of p-values to α

Work in Groups

A researcher collected a sample of 12 fish and measured gill length. They calculated $\bar{X} = 3.12$ and $s = 0.7$. Test the following hypothesis at the $\alpha = 0.05$ level:

$$H_0 : \mu = 3.5$$

$$H_1 : \mu < 3.5$$

Example

Observe sample of 27 newborns in a hospital near Chernobyl and record number of fingers. You calculate $\bar{X} = 9.2$ and $s = 1.8$. You want to test to the following hypothesis at the $\alpha = 0.10$ level:

$$H_0 : \mu = 10$$

$$H_1 : \mu \neq 10$$

Rejection Region from t-Table

Lookup the critical value t^*, based off the degrees of freedom and level of significance, α

Right tailed test:

- Reject H_0 when $t > t^*$

Left tailed test:

- Reject H_0 when $t < -t^*$

Can work back out the rejection region to find range for $\bar{X}$

$$ \daisy{t} = \frac{\bar{X}-\mu}{s/\sqrt{n}} > t^ \implies \bar{X} > t^{} \frac{s}{\sqrt{n}} + \mu $$

Example

Suppose you collect a sample of 17 family incomes. You calculate (s = 6,000). You want to test the following hypothesis:

$$H_0 : \mu = \$29,000$$

$$H_1 : \mu > \$29,000$$

If we test at the $\alpha = 0.05$ level, what is our rejection region?

Example

Suppose you collect a sample of 25 students and record their test scores. You calculate $s = 4.2$. If we're testing the following hypothesis:

$$H_0 : \mu = 82$$

$$H_1 : \mu < 82$$

If we're testing at the $\alpha = 0.01$, when do we reject the null hypothesis?

Example

Back to our newborn example. Observe sample of 27 newborns in a hospital near Chernobyl and record number of fingers. You calculate $s = 1.8$. If we're testing the following hypothesis at the $\alpha = 0.05$ level, what is the rejection region?

$$H_0 : \mu = 10$$

$$H_1 : \mu \neq 10$$

Matched Pairs

Often we are interested in seeing if a particular treatment has an effect

- Does pollution impact your exam scores?
- Does access to birth control increase female labor force participation rates?
- Does taking turmeric improve your focus?

These questions can be solved using a matched pair design

- a matched pair design involves making two observations on the same individual, or one observation on two similar individuals.

Matched Pairs

If the conditions for inference are met, we can use a one-sample t procedures to perform inference about the *mean difference*

Essentially, you just consider an observation as the difference between the pair:

- Example: X_i = (Country's labor force before birth-control $-$ Country's labor force after birth-control)

In these types of problems:

- the parameter interest, μ, is the true population difference
- s is the sample standard deviations of within pair difference

Matched Pairs Example

Counselor selects 25 students who enrolled at the same school in seventh and eighth grade to see if their GPAs improved compared to the previous school year, after the school decided to implement mandatory study halls.

VARIABLE	SAMPLE MEAN	STANDARD DEVIATION
Seventh grade GPA	$\bar{x}_7 = 2.89433$	$s_7 = 0.51053$
Eighth grade GPA	$\bar{x}_8 = 2.96493$	$s_8 = 0.57332$
Difference (8th - 7th)	$\bar{x} = 0.07060$	$s = 0.23495$

She conducts the following hypothesis at $\alpha = 0.05$:

$$H_0 : \mu = 0$$

$$H_1 : \mu > 0,$$

where μ is the mean change in GPA from 7th to 8th grade.

Matched Pairs Example

Calculate p-value of the hypothesis test:

Calculate t-stat:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{0.07 - 0}{.235/\sqrt{25}} = 1.49$$

Calculate range of p-values:

- Row 24, $t = 1.49$ between 0.10 and 0.05 $\implies 0.05 < p < 0.10.$
- $p > \alpha$ so she would not reject the null that the mandatory study halls had no effect on grades

Matched Pairs Example

Say 16 cows are randomly fed two different diets. Milk production was measured for each of the cows during each of the two diets, so that 32 milk production values were recorded. You find the average within cow difference is 0.7 liters of milk, and the sample standard deviation of the within cow difference is 1 liter.

Consider the following hypothesis test:

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0$$

Do we reject the null that the cow's produce the same amount of milk at the $\alpha = 0.05$ significance level?

Matched Pairs Example

Calculate the p-value of hypothesis test:

$$t = \frac{0.7 - 0}{1/\sqrt{16}} = 2.8$$

$t_{15} = 2.6$ when $\alpha = 0.01$ and $t_{15} = 2.95$ when $\alpha = 0.005$

This means the p-value is in between $2 \cdot (0.005)$ and $2 \cdot (0.01)$ because we are looking at a two-tailed test.

Therefore, $0.01 < p\text{-value} < 0.02$,

and we reject the null of no difference because our p-value range is less than $\alpha = 0.05$

Clicker Question -- Midterm Example

Which of the following is an example of a matched pairs design?

a. A teacher compares the pretest and posttest scores of students.
b. A teacher compares the scores of students who had a computer-based method of instruction with the scores of other students who had a traditional method of instruction.
c. A teacher compares the scores of students in her class on a standardized test with the national average score.
d. A teacher calculates the average of students' scores on a pair of tests and wishes to see if this average is larger than 80\%.

When to Use Which Distribution

If we have SRS, Normal X, and σ known:

- Use Z distribution

If we have SRS, Normal X, and σ unknown:

- Use the t distribution

What if we don't know that X is normal?

- $n < 15 \implies$ only use t if X looks very normally distributed.
- $15 < n < 31 \implies$ use t as long as there are no extreme outliers
- $n > 31$, probably okay using Z

Recall Law of Large numbers that says as $n \to \infty$:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to^d Z \sim (0, 1)$$