Elementary Statistics – Inference for Numerical Data Pt. 1

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Kaitlyn Cook (https://www.smith.edu/people/kaitlyn-cook)

Plan for Today

- Inference for numerical variables
  - Single mean
  - Independent means

Inference for Categorical Variables

Pieces of a Hypothesis test:

1. ***Null and Alternative Hypotheses***
2. ***Test Statistic***
3. ***Null Distribution***
4. ***P-value***

Chi-square statistic:

$$E_{ij} = \frac{row\ i\ total \times column\ j\ total}{table\ total}$$

$$X^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Warm Up: Inference for Categorical Variables

Practice

In order to assess whether habitat conditions are related to the sunlight choices a lizard makes for resting, Western fence lizards (Sceloporus occidentalis) were observed across three different microhabitats. Perform a hypothesis test with a significance level of 5% to determine if habitat conditions and sunlight choices are related.

	sunlight			
site	sun	partial	shade	Total
desert	16	32	71	119
mountain	56	36	15	107
valley	42	40	24	106
Total	114	108	110	332

Inference for a Single Mean

So far, we have looked at inference for categorical variables.

Now, we will look at inference for numerical variables, starting with **inference for a single mean**.

Inference for a Single Mean

Our sample statistic ($\bar{x}$) represents our best guess for the true population parameter, (μ). We know this best guess is not perfect; we expect error (variability) due to the sampling process.

Because we can't know the truth directly, we infer the truth via:

1. A confidence interval
2. A hypothesis test

Inference for a Single Mean

In either case, we need the sampling distribution for $\bar{x}$.

We can approximate it via the central limit theorem as long as:

1. The sample's observations are **independent**
2. The sample size is **large enough**, $n \geq 30$, or clearly normally distributed with no outliers

When these conditions are met, variability of $\bar{x}$ is well described by:

$$SE(\bar{x}) = \frac{best\ guess\ of\ \sigma}{\sqrt{n}}$$

Inference for a Single Mean

In either case, we need the sampling distribution for $\bar{x}$.

We can approximate it via the central limit theorem as long as:

1. The sample's observations are **independent**
2. The sample size is **large enough**, $n \geq 30$, or clearly normally distributed with no outliers

When these conditions are met, variability of $\bar{x}$ is well described by:

$$SE(\bar{x}) = \frac{best\ guess\ of\ \sigma}{\sqrt{n}}$$

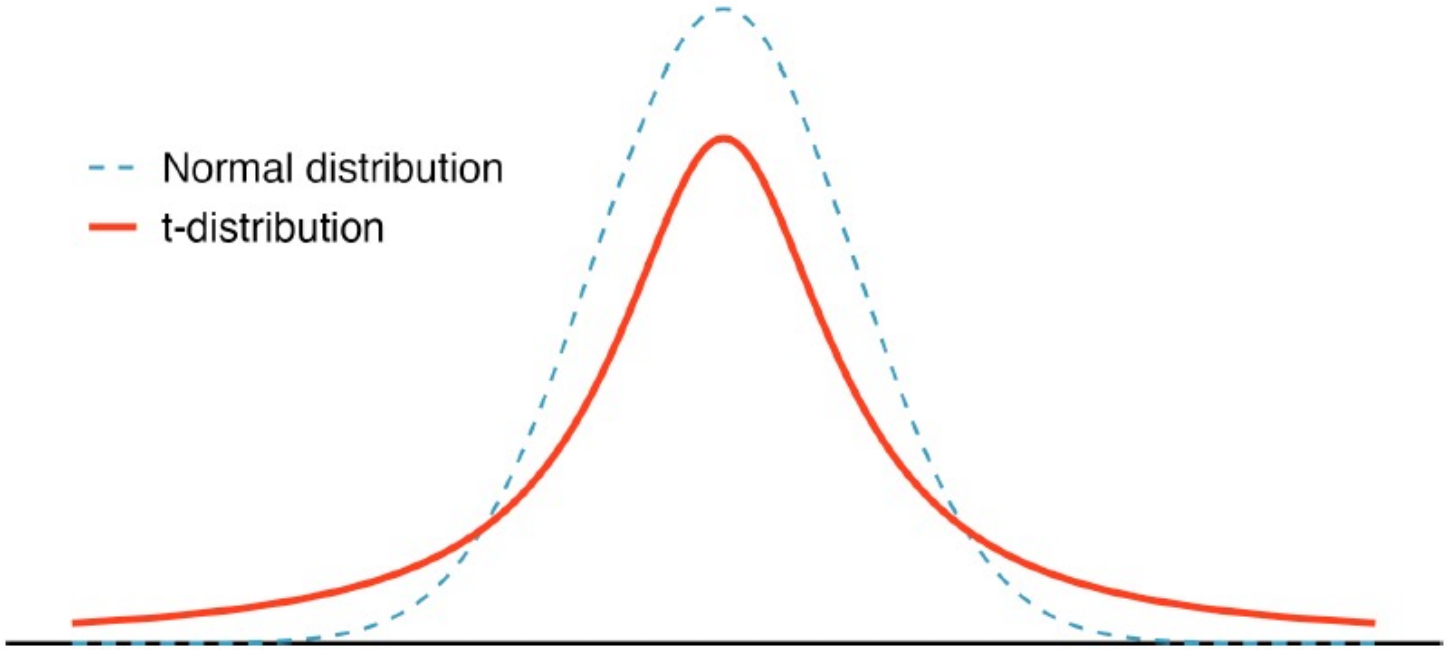
We typically use s (sample variance), as the best guess for σ (population variance). However, this is less precise with small samples.

As a solution, we use the t-distribution to model the sampling distribution of $\bar{x}$

The t-distribution

The t-distribution is similar to the normal distribution, but has thicker tails.

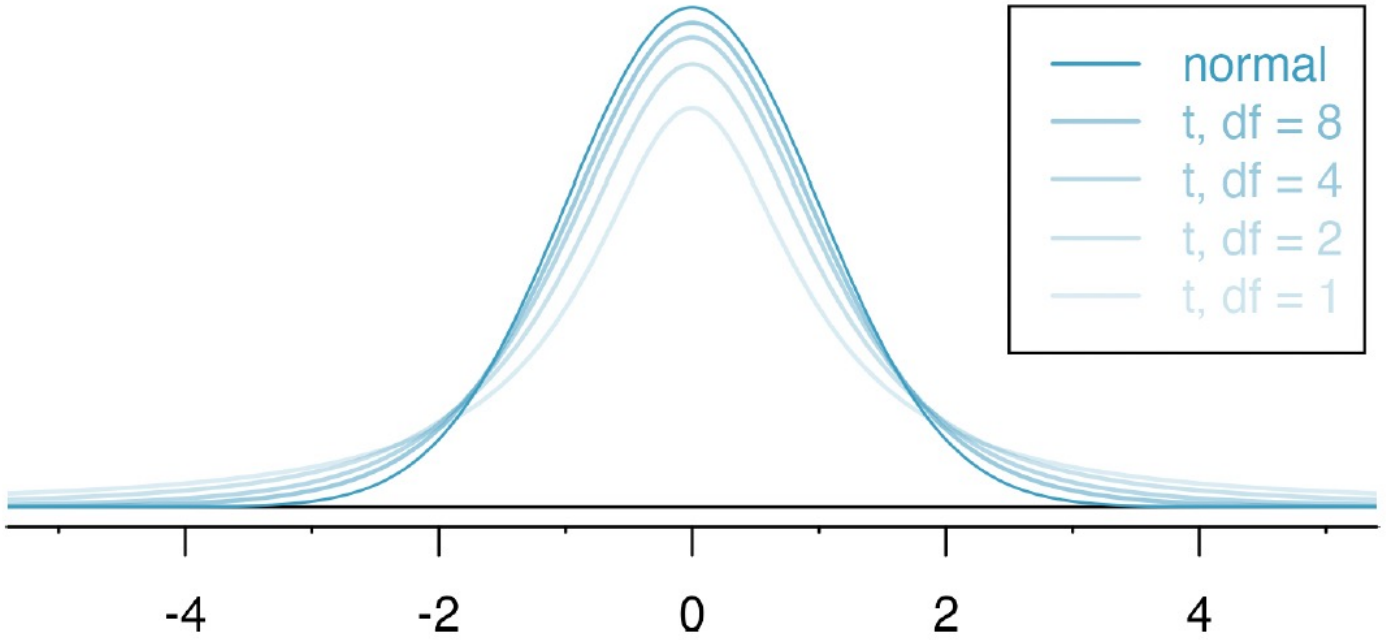
This means observations are more likely to fall more than two standard deviations away from the mean. In other words, it accounts for more variance than the normal distribution.



The t-distribution

Like the Chi-square distribution, the shape of the t-distribution depends on degrees of freedom.

When our sample size is n, we will use a t-distribution with $df = n - 1$ to model the null distribution of the sample mean, $\bar{x}$.



Inference for a Single Mean

Confidence Interval for One Mean

For confidence intervals, we use $\bar{x}$ as the best guess of μ, and s as the best guess of σ, so

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

Inference for a Single Mean

Confidence Interval for One Mean

For confidence intervals, we use $\bar{x}$ as the best guess of μ, and s as the best guess of σ, so

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

We use SE to compute margin of error for our confidence interval: $(\bar{x} - t^*SE, \bar{x} + t^*\mathrm{SE})$

Inference for a Single Mean

Confidence Interval for One Mean

For confidence intervals, we use $\bar{x}$ as the best guess of μ, and s as the best guess of σ, so

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

We use SE to compute margin of error for our confidence interval:
$\left(\bar{x} - t^*_{df} SE, \bar{x} + t^*_{df} \mathrm{SE}\right)$

t^*_{df} is calculated from a specified percentile on the t-distribution with df.

Ex. 5th percentile of a for a 95% confidence

Inference for a Single Mean

Confidence Interval for One Mean

For confidence intervals, we use $\bar{x}$ as the best guess of μ, and s as the best guess of σ, so

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

We use SE to compute margin of error for our confidence interval:
$\left(\bar{x} - t^*_{df} SE, \bar{x} + t^*_{df} \mathrm{SE}\right)$

t^*_{df} is calculated from a specified percentile on the t-distribution with df.

Ex. 5th percentile of a for a 95% confidence

Find t^*_{99} for a 95% CI, a 90% CI, and a 99% CI

Inference for a Single Mean

Confidence Interval for One Mean

For confidence intervals, we use $\bar{x}$ as the best guess of μ, and s as the best guess of σ, so

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

We use SE to compute margin of error for our confidence interval:
$\left(\bar{x} - t^*_{df} SE, \bar{x} + t^*_{df} \mathrm{SE}\right)$

t^*_{df} is calculated from a specified percentile on the t-distribution with df.

Ex. 5th percentile of a for a 95% confidence

You suspect that on average WSU students work 15 hours/week . You perform an experiment to statistically test this suspicion. You sample 100 students and calculate their average number of hours worked per week to be 25, with $s = 5$. Calculate a 95% CI for μ from your $\bar{x}$.

Inference for a Single Mean

Hypothesis Test for One Mean

When the conditions are met so that the distribution of $\bar{x}$ can be modeled with a t-distribution, variability of $\bar{x}$ is well described by:

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

Inference for a Single Mean

Hypothesis Test for One Mean

When the conditions are met so that the distribution of $\bar{x}$ can be modeled with a t-distribution, variability of $\bar{x}$ is well described by:

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

Because we are using the t-distribution, we will need a T-score to find our p-value:

$$T = \frac{\bar{x} - \mu_0}{SE}$$

Degrees of freedom, $df = n - 1$

Inference for a Single Mean

Hypothesis Test for One Mean

When the conditions are met so that the distribution of $\bar{x}$ can be modeled with a t-distribution, variability of $\bar{x}$ is well described by:

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

Because we are using the t-distribution, we will need a T-score to find our p-value:

$$T = \frac{\bar{x} - \mu_0}{SE}$$

Degrees of freedom, $df = n - 1$

You suspect that on average WSU students work 15 hours/week . You perform an experiment to statistically test this suspicion. You sample 100 students and calculate their average number of hours worked per week to be 25 with $s = 5$. Perform a hypothesis test for $H_0: \mu = 15, H_A: \mu \neq 15$. Use $\alpha = 0.05$.

Inference for Two Independent Means

So far, we've done inference to see if our population mean differs from some hypothesized value.

Ex. Is the average number of hours WSU students work per week 15?

Sometimes our research question instead focuses on **comparing means from two independent groups**.

Ex. Is the average number of hours WSU students work per week different than the average number of hours Springfield College students work per week?

Inference for Two Independent Means

So far, we've done inference to see if our population mean differs from some hypothesized value.

> Ex. Is the average number of hours WSU students work per week 15?

Sometimes our research question instead focuses on **comparing means from two independent groups**.

> Ex. Is the average number of hours WSU students work per week different than the average number of hours Springfield College students work per week?

Just like with one mean, we can use sample statistics to infer the population level answer to this question with
(a) a confidence interval and/or
(b) a hypothesis test

Inference for Two Independent Means

To compare two independent means, we look at their difference.

Ex. Is the average number of hours WSU students work per week different than the average number of hours Springfield College students work per week?

To answer this, we need to look at $\bar{x}_{WSU} - \bar{x}_{SC}$

Inference for Two Independent Means

Confidence Interval for Difference Between Two Independent Means

Conditions for $\bar{x}_1 - \bar{x}_2$ to be approximated with the t-distribution:

1. Data are **independent within and between** the two **groups**
2. Each group is has **over 30 observations**, or is clearly normally distributed with no outliers

When the conditions are met so that the distribution of $\bar{x}_1 - \bar{x}_2$ can be modeled by the t-distribution, variability is described by:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{(best\ guess\ for\ \sigma_1)^2}{n_1} + \frac{(best\ guess\ for\ \sigma_2)^2}{n_2}}$$

σ_1, σ_2 are population proportions for each group and n_1, n_2 are the sample sizes for each group

For degrees of freedom, use the smaller of $n_1 - 1, n_2 - 1$

Inference for Two Independent Means

Confidence Interval for Difference Between Two Independent Means

For confidence intervals, we use s_1 and s_2 as the best guess of σ_1 and σ_2, so

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

For degrees of freedom, use the smaller of $n_1 - 1, n_2 - 1$

Inference for Two Independent Means

Confidence Interval for Difference Between Two Independent Means

For confidence intervals, we use s_1 and s_2 as the best guess of σ_1 and σ_2, so

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

For degrees of freedom, use the smaller of $n_1 - 1, n_2 - 1$

We use SE to compute margin of error for our confidence interval:

$$\left((\bar{x}_1 - \bar{x}_2) - t^*_{df} SE, (\bar{x}_1 - \bar{x}_2) + t^*_{df} SE\right)$$

Inference for Two Independent Means

Confidence Interval for Difference Between Two Independent Means

For confidence intervals, we use s_1 and s_2 as the best guess of σ_1 and σ_2, so

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

For degrees of freedom, use the smaller of $n_1 - 1, n_2 - 1$

We use SE to compute margin of error for our confidence interval:

$$\left((\bar{x}_1 - \bar{x}_2) - t^*_{df} SE, (\bar{x}_1 - \bar{x}_2) + t^*_{df} SE\right)$$

You suspect WSU and SC students spend different amounts of time working per week. You perform an experiment to statistically test this suspicion. You sample 120 WSU students and find on average they work 20 hours/week, with $s = 2$. You sample 130 SC students and find on average they work 7 hours/week, with $s = 4$. Calculate a 95% CI for $\mu_{SC} - \mu_{WSU}$ from your $\bar{x}_{SC}$ and $\bar{x}_{WSU}$.

Inference for Two Independent Means

Hypothesis Test for Difference Between Two Independent Means

For a hypothesis test, our null hypothesis will be that there is no difference between means.

Inference for Two Independent Means

Hypothesis Test for Difference Between Two Independent Means

For a hypothesis test, our null hypothesis will be that there is no difference between means.

You suspect WSU and SC students spend different amounts of time working per week. You perform an experiment to statistically test this suspicion. You sample 120 WSU students and find on average they work 20 hours/week, with $s = 2$. You sample 130 SC students and find on average they work 7 hours/week, with $s = 4$.

You want to perform a hypothesis test to see if there is a difference in these means. What is H_0 in terms of $\bar{x}_{SC} - \bar{x}_{WSU}$? What is H_A?

Inference for Two Independent Means

Hypothesis Test for Difference Between Two Independent Means

For a hypothesis test, our null hypothesis will be that there is no difference between means.

We use s_1 and s_2 as the best guess of σ_1 and σ_2, so

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Inference for Two Independent Means

Hypothesis Test for Difference Between Two Independent Means

For a hypothesis test, our null hypothesis will be that there is no difference between means.

We use s_1 and s_2 as the best guess of σ_1 and σ_2, so

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Then, $T = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SE}$, and df is the smaller of $n_1 - 1, n_2 - 1$

Inference for Two Independent Means

Hypothesis Test for Difference Between Two Independent Means

For a hypothesis test, our null hypothesis will be that there is no difference between means.

We use s_1 and s_2 as the best guess of σ_1 and σ_2, so

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Then, $T = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SE}$, and df is the smaller of $n_1 - 1, n_2 - 1$

You suspect WSU and SC students spend different amounts of time working per week. You perform an experiment to statistically test this suspicion. You sample 120 WSU students and find on average they work 20 hours/week, with $s = 2$. You sample 130 SC students and find on average they work 7 hours/week, with $s = 4$.

Finish the hypothesis test. Calculate T, find the p-value, and compare to an α of 0.05.

Hypothesis Test for Difference Between Two Independent Means

Inf
Tw
Inc
Me

A group of researchers who are interested in the possible effects of distracting stimuli during eating, such as an increase or decrease in the amount of food consumption, monitored food intake for a group of 44 patients who were randomized into two equal groups. The treatment group ate lunch while playing solitaire, and the control group ate lunch without any added distractions. Patients in the treatment group ate 52.1 grams of biscuits, with a standard deviation of 45.1 grams, and patients in the control group ate 27.1 grams of biscuits, with a standard deviation of 26.4 grams. Do these data provide convincing evidence that the average food intake (measured in amount of biscuits consumed) is different for the patients in the treatment group compared to the control group? Assume that conditions for conducting inference using mathematical models are satisfied.