Module 2: Hypothesis Testing and Statistical Inference (5 lectures)

Reading: Statistics for Business and Economics, Ch. 5-7
Confidence intervals

Given the sample mean $\bar{x}$ and standard deviation $s$, we want to draw conclusions about the population mean $\mu$, of the form: “There is a 95% chance that $\mu$ is between ____ and ____.”

**Example:** We want to know the average amount of money $\mu$ that a Pittsburgh household spends yearly on Internet purchases. For a random sample of $N = 36$ households, we compute a mean of $350 and standard deviation of $180. What can we conclude about $\mu$?

Intuitively, $\mu$ is likely to be close to $350$, but unlikely to be exactly $350$.

To estimate how close $\mu$ is to $350$, we use the Central Limit Theorem.

If the population has any distribution with mean $\mu$ and standard deviation $\sigma$, and if $N \geq 30$, then the sample mean $\bar{x}$ is normally distributed, with mean $\mu$ and standard deviation $\sigma / \sqrt{N}$.
Large-sample confidence intervals

Given the sample mean $\bar{x}$ and standard deviation $s$, we want to draw conclusions about the population mean $\mu$, of the form: “There is a 95% chance that $\mu$ is between ____ and ____.”

**Step 1:** Using an inverse table lookup, we know that 95% of the area under a normal curve lies between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$.

\[
0.95 = 2 \times F(z_c) \quad \longrightarrow \quad z_c = F^{-1}(0.475) = 1.96
\]

**Step 2:** Assuming $N \geq 30$, we know that $\bar{x}$ is normally distributed with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma / \sqrt{N} \approx s / \sqrt{N}$.

\[
\Pr(\mu_{\bar{x}} - 1.96\sigma_{\bar{x}} \leq \bar{x} \leq \mu_{\bar{x}} + 1.96\sigma_{\bar{x}}) = 0.95
\]

\[
\Pr(\mu - 1.96(s / \sqrt{N}) \leq \bar{x} \leq \mu + 1.96(s / \sqrt{N})) = 0.95
\]

\[
\Pr(\bar{x} - 1.96(s / \sqrt{N}) \leq \mu \leq \bar{x} + 1.96(s / \sqrt{N})) = 0.95
\]

“There is a 95% chance that $\mu$ is between $\bar{x} - 1.96(s / \sqrt{N})$ and $\bar{x} + 1.96(s / \sqrt{N})$.”
Large-sample confidence intervals

**Example:** We want to know the average amount of money \( \mu \) that a Pittsburgh household spends yearly on Internet purchases. For a random sample of \( N = 36 \) households, we compute a mean of $350 and standard deviation of $180. What can we conclude about \( \mu \)?

**Answer:** There is a 95% chance that \( \mu \) is between $291.20 and $408.80.

\[
\bar{x} \pm 1.96 \left( \frac{s}{\sqrt{N}} \right)
\]

\[
350 \pm 1.96(180 / 6)
\]

What if we sampled 10,000 households, and obtained the same \( \bar{x} \) and \( s \)?

**Answer:** There is a 95% chance that \( \mu \) is between $346.47 and $353.53.

\[
\bar{x} \pm 1.96 \left( \frac{s}{\sqrt{N}} \right)
\]

\[
350 \pm 1.96(180 / 100)
\]

“There is a 95% chance that \( \mu \) is between \( \bar{x} - 1.96(s / \sqrt{N}) \) and \( \bar{x} + 1.96(s / \sqrt{N}) \).”
Large-sample confidence intervals

**Example:** We want to know the average amount of money $\mu$ that a Pittsburgh household spends yearly on Internet purchases. For a random sample of $N = 36$ households, we compute a mean of $350$ and standard deviation of $180$. What can we conclude about $\mu$?

**Answer:** There is a 95% chance that $\mu$ is between $291.20$ and $408.80$.

$$350 \pm 1.96\left(\frac{180}{6}\right)$$

How many households must we sample to be 95% certain that $\mu$ is within a range of $5$ of the sample mean?

$$1.96\left(\frac{s}{\sqrt{N}}\right) = 5 \quad \rightarrow \quad N = \left(\frac{1.96s}{5}\right)^2 = 4,979$$

“There is a 95% chance that $\mu$ is between $\bar{x} - 1.96\left(\frac{s}{\sqrt{N}}\right)$ and $\bar{x} + 1.96\left(\frac{s}{\sqrt{N}}\right)$.”
Large-sample confidence intervals

**Example:** We want to know the average amount of money \( \mu \) that a Pittsburgh household spends yearly on Internet purchases. For a random sample of \( N = 36 \) households, we compute a mean of $350 and standard deviation of $180. What can we conclude about \( \mu \)?

What if we would like to be 99% certain of the value of \( \mu \)?

**Step 1:** Using an inverse table lookup, we know that proportion \( c \) of the area under a normal curve lies between \( \mu - z_c \sigma \) and \( \mu + z_c \sigma \), where:

\[
c = 2 \cdot F(z_c) \quad \rightarrow \quad z_c = F^{-1}(c / 2)
\]

\( z_c \) is the confidence threshold corresponding to the confidence interval \( c \).

**Step 2:** Assuming \( N \geq 30 \), we know that \( \bar{x} \) is normally distributed with mean \( \mu_{\bar{x}} = \mu \) and standard deviation \( \sigma_{\bar{x}} = \sigma / \sqrt{N} \approx s / \sqrt{N} \).

“There is a probability \( c \) that \( \mu \) is between \( \bar{x} - z_c(s / \sqrt{N}) \) and \( \bar{x} + z_c(s / \sqrt{N}) \).”
Small-sample confidence intervals

**Example:** We want to know the average amount of money $\mu$ that a Pittsburgh household spends yearly on Internet purchases. For a random sample of $N = 16$ households, we compute a mean of $350$ and standard deviation of $180$. What can we conclude about $\mu$?

When the sample size is small ($N < 30$), we must deal with two problems:

1. The Central Limit Theorem does not guarantee that the sample mean $\bar{x}$ is normally distributed.

2. The sample standard deviation $s$ may not be a good estimate of the population standard deviation $\sigma$.

The first problem means that we can only do small-sample inference when we believe that the population is approximately normally distributed.

Small to moderate deviations from normality are ok, but for highly skewed distributions we must use a non-parametric test (see McClave, Ch. 14).
Small-sample confidence intervals

**Example:** We want to know the average amount of money \( \mu \) that a Pittsburgh household spends yearly on Internet purchases. For a random sample of \( N = 16 \) households, we compute a mean of $350 and standard deviation of $180. What can we conclude about \( \mu \)?

When the sample size is small (\( N < 30 \)), we must deal with two problems:

1. The Central Limit Theorem does not guarantee that the sample mean \( \bar{x} \) is normally distributed.

2. The sample standard deviation \( s \) may not be a good estimate of the population standard deviation \( \sigma \).

The second problem means that we must use a confidence threshold based on the **t-distribution** rather than the normal distribution.

The t-score is calculated and interpreted just like the z-score, but it accounts for the error in using \( s \) to estimate \( \sigma \).
Small-sample confidence intervals

Example: We want to know the average amount of money $\mu$ that a Pittsburgh household spends yearly on Internet purchases. For a random sample of $N = 16$ households, we compute a mean of $350$ and standard deviation of $180$. What can we conclude about $\mu$?

“There is a probability $c$ that $\mu$ is between $\bar{x} - t_c(s / \sqrt{N})$ and $\bar{x} + t_c(s / \sqrt{N})$.”

We obtain the value of $t_c$ from a t-score table, using two values: the confidence interval $c$, and the number of degrees of freedom $df = N - 1$.

For example, for a 95% confidence interval, and $N - 1 = 15$ degrees of freedom, we have $t_c = 2.131$ (instead of $z_c = 1.96$).

In our example, there is a 95% probability that $\mu$ is between $254.10$ and $445.90$.

For small samples, the uncertainty about $\sigma$ leads to a wider range for $\mu$. 

$$350 \pm 2.131(180 / 4)$$
Understanding confidence intervals

“There is a 95% probability that $\mu$ is between $250$ and $300$.”

$\mu$ is a fixed quantity that we are trying to estimate.

To do so, we choose $N$ samples from the population and calculate their sample mean $\bar{x}$ and standard deviation $s$.

This lets us calculate a 95% confidence interval from $\bar{x} - \varepsilon$ to $\bar{x} + \varepsilon$.

For large samples, $\varepsilon = 1.96s / \sqrt{N}$. 
Understanding confidence intervals

“There is a 95% probability that $\mu$ is between $250$ and $300$.”

$\mu$ is a fixed quantity that we are trying to estimate.

To do so, we choose $N$ samples from the population and calculate their sample mean $\bar{x}$ and standard deviation $s$.

This lets us calculate a 95% confidence interval from $\bar{x} - \varepsilon$ to $\bar{x} + \varepsilon$.

For large samples, $\varepsilon = 1.96s / \sqrt{N}$.

If we had chosen a different sample, we would have calculated a different confidence interval.
Understanding confidence intervals

“There is a 95% probability that $\mu$ is between $250 and $300.”

$\mu$ is a fixed quantity that we are trying to estimate.

To do so, we choose N samples from the population and calculate their sample mean $\bar{x}$ and standard deviation s.

This lets us calculate a 95% confidence interval from $\bar{x} - \varepsilon$ to $\bar{x} + \varepsilon$.

For large samples, $\varepsilon = 1.96s / \sqrt{N}$.

If we had chosen a different sample, we would have calculated a different confidence interval.

95% of the time, the interval will contain $\mu$, and 5% of the time it will not.
We must carefully consider the tradeoffs between the number of samples N, the length of the interval $2\varepsilon$, and the confidence level $c$.

For a fixed number of samples N, a higher confidence level means a larger interval for $\mu$. For example, you may be 90% certain that $\mu$ is between 250 and 350, and 99.9% certain that $\mu$ is between 200 and 400.
Understanding confidence intervals

We must carefully consider the tradeoffs between the number of samples $N$, the length of the interval $2\varepsilon$, and the confidence level $c$.

For a fixed number of samples $N$, a higher confidence level means a larger interval for $\mu$. For example, you may be 90% certain that $\mu$ is between 250 and 350, and 99.9% certain that $\mu$ is between 200 and 400.

If we increase the number of samples, we can do one of two things:
1. Keep the confidence level constant, and shorten the interval.
   “90% certain that $\mu$ is between 275 and 325.”
2. Keep the interval length constant, and increase the confidence.
   “99.9% certain that $\mu$ is between 250 and 350.”

Disadvantage: taking more samples may be expensive or infeasible.
Confidence intervals for proportions

In a random sample of 400 U.S. college students, 40% were in favor of the president’s domestic policy decisions, and 60% opposed. What can we conclude about \( p \), the proportion of students that support the president’s domestic policy?

Define \( x_i = 1 \) if the \( i \)th individual supports the president’s policy, and \( x_i = 0 \) if the individual opposes his policy.

Then the sample mean \( \overline{x} = \sum x_i / N \) is the proportion of the sampled individuals supporting the president’s policy.

According to the Central Limit Theorem, \( \overline{x} \) will be normally distributed for \( N \geq 30 \), with mean \( p \) and std. dev. \( \sqrt{p(1-p)} / N \).

\[
\Pr(p - 1.96\sqrt{p(1-p)} / N \leq \overline{x} \leq p + 1.96\sqrt{p(1-p)} / N ) = 0.95
\]

\[
\Pr(\overline{x} - 1.96\sqrt{\overline{x}(1-\overline{x})} / N \leq p \leq \overline{x} + 1.96\sqrt{\overline{x}(1-\overline{x})} / N ) = 0.95
\]

\[
\Pr(\overline{x} - 1.96\sqrt{\overline{x}(1-\overline{x})} / N \leq \overline{x} \leq \overline{x} + 1.96\sqrt{\overline{x}(1-\overline{x})} / N ) = 0.95
\]
Confidence intervals for proportions

In a random sample of 400 U.S. college students, 40% were in favor of the president's domestic policy decisions, and 60% opposed. What can we conclude about \( p \), the proportion of students that support the president's domestic policy?

\[
\mu = p \\
\sigma = \sqrt{p(1-p)}
\]

Define \( x_i = 1 \) if the \( i \)th individual supports the president's policy, and \( x_i = 0 \) if the individual opposes his policy.

Then the sample mean \( \bar{x} = \sum x_i / N \) is the proportion of the sampled individuals supporting the president's policy.

According to the Central Limit Theorem, \( \bar{x} \) will be normally distributed for \( N \geq 30 \), with mean \( p \) and std. dev. \( \sqrt{p(1-p)} / N \).

“There is a probability \( c \) that \( p \) is between \( \bar{x} - z_c(\sqrt{\bar{x}(1-\bar{x})} / N) \) and \( \bar{x} + z_c(\sqrt{\bar{x}(1-\bar{x})} / N) \).”

Note: this method is not accurate for very small or very large \( \bar{x} \). See McClave, Section 5.4, for more details.
Confidence intervals for proportions

In a random sample of 400 U.S. college students, 40% were in favor of the president’s domestic policy decisions, and 60% opposed. What can we conclude about p, the proportion of students that support the president’s domestic policy?

Answer: There is a 95% probability that p is between 0.352 and 0.448.

“40% of students support the president’s policy, with a sampling error of +/- 4.8%.”

0.4 ± 1.96(√(0.4)(0.6) / 400)

How many samples would we need to estimate p within +/- 1%?

1.96 √(0.4)(0.6) / N = 0.01 → N = (0.4)(0.6)(1.96 / 0.01)^2 = 9,220

“There is a probability c that p is between \( \bar{x} - z_c(\sqrt{\bar{x}(1-\bar{x}) / N}) \) and \( \bar{x} + z_c(\sqrt{\bar{x}(1-\bar{x}) / N}) \).”
Confidence intervals for proportions

In a random sample of 400 U.S. college students, 40% were in favor of the president’s domestic policy decisions, and 60% opposed. What can we conclude about p, the proportion of students that support the president’s domestic policy?

Answer: There is a 95% probability that p is between 0.352 and 0.448.

“40% of students support the president’s policy, with a sampling error of +/- 4.8%.”

How many samples would we need to estimate p within +/- 1%? (if we didn’t know in advance that \( \bar{x} = 0.4 \))

Use a conservative bound: \( \bar{x}(1-\bar{x}) \) is maximized at \( \bar{x} = 0.5 \).

\[
1.96 \sqrt{(0.5)(0.5)} / N = 0.01 \quad \rightarrow \quad N = (0.5)(0.5)(1.96 / 0.01)^2 = 9,604
\]
Hypothesis testing

We have been drawing inferences about $\mu$ using confidence intervals:
“There is a 95% chance that $\mu$ is between ____ and ____.”

What if we want to test a specific claim about the value of $\mu$?

In each case, we want to decide which of two possible hypotheses is true:

\[ H_1 : \mu > 40,000 \quad H_1 : \mu \neq \mu_0 \]
\[ H_0 : \mu \leq 40,000 \quad H_0 : \mu = \mu_0 \]

where $\mu$ is some objective measure of productivity, and $\mu_0$ is its historical average.

“Is the mean income $\mu$ of Pittsburgh steelworkers over $40,000?$”

Does our new integrated development environment affect programmer productivity?
We want to test the alternative hypothesis $H_1 : \mu \neq 1000$ against the null hypothesis $H_0 : \mu = 1000$.

Let us assume that we want to measure productivity in terms of lines of production-quality code written, and that historically we have achieved an average of $\mu_0 = 1000$ lines of code per day.

Generally, the alternative hypothesis $H_1$ indicates that there is an effect (e.g. significant increase or decrease in some quantity) while the null hypothesis $H_0$ indicates that there is no effect (e.g. the quantity has not changed significantly).

Our test will give one of two possible outcomes:

1. We can reject the null hypothesis, and thus the alternative hypothesis is true.
2. We cannot reject the null hypothesis. This does not necessarily mean that the null is true!

— “We can conclude that $\mu \neq 1000$.”
— “We do not have sufficient evidence to conclude that $\mu \neq 1000$."

Hypothesis testing
Hypothesis testing

We want to test the alternative hypothesis $H_1 : \mu \neq 1000$ against the null hypothesis $H_0 : \mu = 1000$.

Let us assume that we want to measure productivity in terms of lines of production-quality code written, and that historically we have achieved an average of $\mu_0 = 1000$ lines of code per day.

Key idea: the sample evidence must strongly contradict the null hypothesis for us to reject it in favor of the alternative.

Our test will give one of two possible outcomes:

1. We can reject the null hypothesis, and thus the alternative hypothesis is true.
2. We cannot reject the null hypothesis. This does not necessarily mean that the null is true!

---

“We can conclude that $\mu \neq 1000$.”

“We do not have sufficient evidence to conclude that $\mu \neq 1000$.”
Identifying $H_1$ and $H_0$

A statistical hypothesis is an assumption about some parameter of a population, such as the population mean $\mu$ or population proportion $p$.

The alternative hypothesis $H_1$ is some claim about a parameter that you want to demonstrate.

The null hypothesis $H_0$ is the assumption about this parameter that you must reject in order to show that $H_1$ is true.

“Support for the new billing system is less than 50%.”

“The community’s average yearly expenditure on computing supplies is greater than $40$.”

“The company’s charitable giving rate did not equal the historical mean of 0.2% of net equity.”
Two-sided hypothesis tests

We want to test $H_1: \mu \neq \mu_0$ against $H_0: \mu = \mu_0$.

**Solution:** use the sample mean $\bar{x}$ and sample standard deviation $s$, and reject the null hypothesis if $\bar{x}$ is sufficiently far from $\mu_0$.

Assume $\mu = \mu_0$. Then if $N \geq 30$, $\bar{x}$ is normally distributed with mean $\mu_0$ and standard deviation $\sigma / \sqrt{N} \approx s / \sqrt{N}$. Thus the z-score of $\bar{x}$ is $z = (\bar{x} - \mu_0) / (s / \sqrt{N})$.

Reject $H_0$ if $z < -z_c$ or $z > z_c$.

Typically we use $z_c = 1.96$, corresponding to a **significance level** of $\alpha = 0.05$.

Reject $H_0$ if $\bar{x} < \mu_0 - \varepsilon$.

Do not reject $H_0$ if $\mu_0 - \varepsilon \leq \bar{x} \leq \mu_0 + \varepsilon$.

Reject $H_0$ if $\bar{x} > \mu_0 + \varepsilon$. 
Two-sided hypothesis tests

Based on historical data, our team of programmers produces an average of 1000 lines of production-quality code per day. In the last 36 days, our team has used a new integrated development environment, producing a mean of 1100 lines of production-quality code and standard deviation of 300 lines. Can we conclude that the new environment affects programmer productivity?

\[ H_1 : \mu \neq 1000 \]
\[ H_0 : \mu = 1000 \]

If \( H_0 \) was true, \( \bar{x} \) would be normally distributed with mean 1000 and standard deviation \( \frac{300}{\sqrt{36}} = 50 \).

The z-score corresponding to \( \bar{x} = 1100 \) is \( z = (1100 - 1000) / 50 = 2 \).

We can reject \( H_0 \) since \( z < -1.96 \) or \( z > 1.96 \).

Assuming a significance level of \( \alpha = 0.05 \) and the corresponding threshold \( z_c = 1.96 \) for a two-sided test, we can reject the null hypothesis and conclude that \( \mu \neq 1000 \). The new environment does affect productivity!
One-sided hypothesis tests

We want to test $H_1: \mu > \mu_0$ against $H_0: \mu \leq \mu_0$.

**Solution:** use the sample mean $\bar{x}$ and sample standard deviation $s$, and reject the null hypothesis if $\bar{x}$ is sufficiently higher than $\mu_0$.

Assume $\mu = \mu_0$. Then if $N \geq 30$, $\bar{x}$ is normally distributed with mean $\mu_0$ and standard deviation $\sigma / \sqrt{N} \approx s / \sqrt{N}$. Thus the z-score of $\bar{x}$ is $z = (\bar{x} - \mu_0) / (s / \sqrt{N})$.

Reject $H_0$ if $z > z_c$.

Typically we use $z_c = 1.645$, corresponding to a significance level of $\alpha = 0.05$.

Do not reject $H_0$ if $\bar{x} \leq \mu_0 + \varepsilon$.

Reject $H_0$ if $\bar{x} > \mu_0 + \varepsilon$. 
A computer supplies retail chain has a policy of only opening stores in communities where households spend more than $40 per year on computing supplies and equipment. A survey of 100 households in Monroeville finds that average expenditures in the sample are $40.50 with a standard deviation of $10. Is this strong evidence that the community spends more than $40?

\[
H_1 : \mu > 40 \\
H_0 : \mu \leq 40
\]

If \( H_0 \) was true with \( \mu = 40 \), \( \bar{x} \) would be normally distributed with mean 40 and standard deviation \( \frac{10}{\sqrt{100}} = 1 \).

The z-score corresponding to \( \bar{x} = 40.50 \) is \( z = \frac{40.50 - 40}{1} = 0.5 \).

We cannot reject \( H_0 \) since \( z \leq 1.645 \).

Assuming a significance level of \( \alpha = 0.05 \) and the corresponding threshold \( z_c = 1.645 \) for a one-sided test, we cannot reject the null hypothesis. We do not have sufficient evidence to conclude that the community spends more than $40.
One-sided hypothesis tests

We want to test $H_1: \mu < \mu_0$ against $H_0: \mu \geq \mu_0$.

**Solution:** use the sample mean $\bar{x}$ and sample standard deviation $s$, and reject the null hypothesis if $\bar{x}$ is sufficiently lower than $\mu_0$.

Assume $\mu = \mu_0$. Then if $N \geq 30$, $\bar{x}$ is normally distributed with mean $\mu_0$ and standard deviation $\sigma / \sqrt{N} \approx s / \sqrt{N}$. Thus the z-score of $\bar{x}$ is $z = (\bar{x} - \mu_0) / (s / \sqrt{N})$.

Reject $H_0$ if $z < -z_c$.

Typically we use $z_c = 1.645$, corresponding to a significance level of $\alpha = 0.05$.

Reject $H_0$ if $\bar{x} < \mu_0 - \varepsilon$.

Do not reject $H_0$ if $\bar{x} \geq \mu_0 - \varepsilon$. 
One-sided hypothesis tests

A survey of 400 customers shows that 43% prefer the new on-line bill payment system to the old pay-by-mail system. Is this sufficient evidence to show that a majority of customers do not prefer the new system?

\[ \begin{align*}
H_1 : & \quad p < 0.5 \\
H_0 : & \quad p \geq 0.5
\end{align*} \]

If \( H_0 \) was true with \( p = 0.5 \), \( \bar{x} \) would be normally distributed with mean 0.5 and standard deviation \( \sqrt{(0.5)(0.5) / 400} = .025 \).

The z-score corresponding to \( \bar{x} = 0.43 \) is \( z = (0.43 - 0.5) / 0.025 = -2.8 \).

We can reject \( H_0 \) since \( z < -1.645 \).

Assuming a significance level of \( \alpha = 0.05 \) and the corresponding threshold \( z_c = 1.645 \) for a one-sided test, we can reject the null hypothesis and conclude that \( p < 0.5 \). A majority of customers do not prefer the new system!
Based on historical data, our team of programmers produces an average of 1000 lines of production-quality code per day. In the last 16 days, our team has used a new integrated development environment, producing a mean of 1100 lines of production-quality code and standard deviation of 300 lines. Can we conclude that the new environment affects programmer productivity?

\[ H_1 : \mu \neq 1000 \]
\[ H_0 : \mu = 1000 \]

If \( H_0 \) was true, \( \bar{x} \) would follow a t-distribution with mean 1000, standard deviation \( \frac{300}{\sqrt{16}} = 75 \), and \( 16 - 1 = 15 \) degrees of freedom.

The t-value threshold corresponding to \( \alpha = 0.05 \) and 15 dof is \( t_c = 2.131 \).

The t-score corresponding to \( \bar{x} = 1100 \) is \( t = \frac{(1100 - 1000)}{75} = 1.33 \).

We cannot reject \( H_0 \) since \(-2.131 \leq t \leq 2.131\).

Assuming a significance level of \( \alpha = 0.05 \) and the corresponding threshold \( t_c = 2.131 \) for a two-sided test with 15 degrees of freedom, we do not have sufficient evidence to conclude that the new environment affects productivity.
Review of hypothesis tests

We want to compare $H_1$: “there is an effect” vs. $H_0$: “there is no effect.”

$H_1$: $\mu > \mu_0$, $\mu < \mu_0$, $\mu \neq \mu_0$

$H_0$: $\mu \leq \mu_0$, $\mu \geq \mu_0$, $\mu = \mu_0$ ← $H_0$ always contains $\mu = \mu_0$.

Step 1: Find how the observation $\bar{x}$ would be distributed if $H_0$: $\mu = \mu_0$.

Large samples: $\text{Normal}(\mu_0, s / \sqrt{N})$.
Small samples: $\text{t-dist}(\mu_0, s / \sqrt{N}, N – 1 \text{ dof})$

Step 2: If $\bar{x}$ is far enough from $\mu_0$ in the desired direction(s), reject $H_0$.

Large samples:
For $\mu > \mu_0$: reject $H_0$ when $\bar{x} > \mu_0 + z_c (s / \sqrt{N})$, i.e. when $z > z_c$.
For $\mu < \mu_0$: reject $H_0$ when $\bar{x} < \mu_0 - z_c (s / \sqrt{N})$, i.e. when $z < -z_c$.
For $\mu \neq \mu_0$: reject $H_0$ when $| \bar{x} - \mu_0 | > z_c (s / \sqrt{N})$, i.e. when $| z | > z_c$.

Small samples: same except use $t$ and $t_c$ instead of $z$ and $z_c$.

How to choose our threshold $z_c$ or $t_c$?
Review of hypothesis tests

We want to compare $H_1$: “there is an effect” vs. $H_0$: “there is no effect.”

$H_1$: $p > p_0$, $p < p_0$, $p \neq p_0$
$H_0$: $p \leq p_0$, $p \geq p_0$, $p = p_0$  ← $H_0$ always contains $p = p_0$.

**Step 1:** Find how the observation $\bar{x}$ would be distributed if $H_0$: $p = p_0$.

Large samples: Normal($p_0$, $\sqrt{p_0(1-p_0)/N}$).

**Step 2:** If $\bar{x}$ is far enough from $p_0$ in the desired direction(s), reject $H_0$.

For $p > p_0$: reject $H_0$ when $\bar{x} > p_0 + z_c \sqrt{p_0(1-p_0)/N}$, i.e. when $z > z_c$.
For $p < p_0$: reject $H_0$ when $\bar{x} < p_0 - z_c \sqrt{p_0(1-p_0)/N}$, i.e. when $z < -z_c$.
For $p \neq p_0$: reject $H_0$ when $|\bar{x} - p_0| > z_c \sqrt{p_0(1-p_0)/N}$, i.e. when $|z| > z_c$.

How to choose our threshold $z_c$?
Significance levels

The significance level $\alpha$ is the probability of incorrectly rejecting the null hypothesis $H_0$, if the null is true.

If the null is true and $N$ is large, the z-score $z = (\bar{x} - \mu_0) / (s / \sqrt{N})$ will be normally distributed with mean 0 and standard deviation 1.

Probability of incorrectly rejecting the null: $\alpha = 1 - 2 \cdot F(z_c)$ for a 2-sided test.

Let us assume that we are willing to accept a 5% probability of incorrectly rejecting the null ($\alpha = .05$).

Then $z_c = F^{-1}(\frac{1 - \alpha}{2}) = F^{-1}(0.475) = 1.96$.

Same as 95% confidence interval!

Reject $H_0$ if $z < -z_c$.

Reject $H_0$ if $z > z_c$. 

- $z_c$ 
  0 
  $z_c$
Significance levels

The significance level $\alpha$ is the probability of incorrectly rejecting the null hypothesis $H_0$, if the null is true.

If the null is true and $N$ is large, the z-score $z = (\bar{x} - \mu_0) / (s / \sqrt{N})$ will be normally distributed with mean 0 and standard deviation 1.

Probability of incorrectly rejecting the null: $\alpha = 0.5 - F(z_c)$ for a 1-sided test.

Let us assume that we are willing to accept a 5% probability of incorrectly rejecting the null ($\alpha = .05$).

Then $z_c = F^{-1}(0.5 - \alpha) = F^{-1}(0.45) = 1.645$. Reject $H_0$ if $z > z_c$.
Significance levels

The significance level $\alpha$ is the probability of incorrectly rejecting the null hypothesis $H_0$, if the null is true.

If the null is true and $N$ is small, the t-score $t = (\bar{x} - \mu_0) / (s / \sqrt{N})$ will be t-distributed with mean 0, standard deviation 1, and $N - 1$ degrees of freedom.

To find the t-score threshold $t_c$ for a 1-sided test, for a given significance level $\alpha$:
Look up the value $t_\alpha$ with $N - 1$ degrees of freedom using the t-score table.

To find the t-score threshold $t_c$ for a 2-sided test, for a given significance level $\alpha$:
Look up the value $t_{\alpha/2}$ with $N - 1$ degrees of freedom using the t-score table.

Just as for confidence intervals, the t-score threshold $t_c$ will be larger than the corresponding z-score threshold $z_c$, to account for the uncertainty in using the sample standard deviation $s$ to estimate the population standard deviation $\sigma$.

This means that it is harder to reject the null hypothesis when $N$ is small.
Type I and Type II errors

Key idea: Making inferences about the population parameters based on sample statistics is inherently uncertain and thus subject to error.

Our decision

<table>
<thead>
<tr>
<th>Decision</th>
<th>If $H_0$ is true</th>
<th>If $H_0$ is false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do not reject $H_0$</td>
<td>CORRECT</td>
<td>TYPE II ERROR</td>
</tr>
<tr>
<td>Reject $H_0$</td>
<td>TYPE I ERROR</td>
<td>CORRECT</td>
</tr>
</tbody>
</table>

Type I

$$
\mu_0 \quad \bar{x}
$$

(TRUE)

Type II

$$
\mu_0 \quad \bar{x}
$$

(FALSE)
Type I and Type II errors

Key idea: Making inferences about the population parameters based on sample statistics is inherently uncertain and thus subject to error.

Let $\alpha$ = probability of making a type I error (rejecting a true null)
Let $\beta$ = probability of making a type II error (failing to reject a false null)

As discussed previously, $\alpha$ is the total probability in the tails of the null distribution.
$\beta$ is hard to calculate: it depends on how far the true mean $\mu$ is from $\mu_0$. 
Type I and Type II errors

Key idea: Making inferences about the population parameters based on sample statistics is inherently uncertain and thus subject to error.

Let \( \alpha \) = probability of making a type I error (rejecting a true null)
Let \( \beta \) = probability of making a type II error (failing to reject a false null)

As discussed previously, \( \alpha \) is the total probability in the tails of the null distribution.
\( \beta \) is hard to calculate: it depends on how far the true mean \( \mu \) is from \( \mu_0 \).

Increasing the width of the interval decreases \( \alpha \) but increases \( \beta \).
Balancing Type I and Type II errors

Step 1: List and quantify the costs of a type I error.
Step 2: List and quantify the costs of a type II error.
Step 3: Estimate the distance of the true mean $\mu$ from $\mu_0$.
   (How large of an effect do we expect to see?)
Step 4: Ask an expert to calculate the tradeoff between $\alpha$ and $\beta$.
Step 5: Choose a value of $\alpha$ that reasonably balances these costs.

A computer supplies retail chain has a policy of only opening stores in communities where households spend more than $40 per year on computing supplies and equipment. A survey of 100 households in Monroeville finds that average expenditures in the sample are $40.50 with a standard deviation of $10. Is this strong evidence that the community spends more than $40?

What is a Type I error, and what are its consequences?

What is a Type II error, and what are its consequences?

How much do we expect the communities we are interested in to spend?
Balancing Type I and Type II errors

Step 1: List and quantify the costs of a type I error.
Step 2: List and quantify the costs of a type II error.
Step 3: Estimate the distance of the true mean $\mu$ from $\mu_0$.
   (How large of an effect do we expect to see?)
Step 4: Ask an expert to calculate the tradeoff between $\alpha$ and $\beta$.
Step 5: Choose a value of $\alpha$ that reasonably balances these costs.

A computer supplies retail chain has a policy of only opening stores in communities where households spend more than $40 per year on computing supplies and equipment. A survey of 100 households in Monroeville finds that average expenditures in the sample are $40.50 with a standard deviation of $10. Is this strong evidence that the community spends more than $40?

**Option I**: We will open stores in 80% of communities that spend $50 or more ($\beta = 0.2$ for $\mu = 50$) but also in 10% of communities that spend $40 or less ($\alpha = 0.1$).

**Option II**: We will open stores in 50% of communities that spend $50 or more ($\beta = 0.5$ for $\mu = 50$) but also in 5% of communities that spend $40 or less ($\alpha = 0.05$).

Not happy with any of these options? Then collect more samples!
Using p-values for hypothesis testing

We have learned one way to do hypothesis testing:
1. Choose a significance level $\alpha$.
2. Compute the corresponding z-score threshold $z_c$.
3. Compare the observed z-score $z$ to the threshold $z_c$.
4. Reject $H_0$ if the z-score falls outside the threshold.

Question: What if we don’t know a good value for $\alpha$?

Answer: report the observed significance level, or p-value, from your test.

For example, a p-value of 0.04 would mean, “Reject the null if your chosen significance level $\alpha$ is higher than 0.04.”

Someone else can then choose whether or not to reject the null, based on the value of $\alpha$ that they think is reasonable.

A lower p-value means that the data disagrees more strongly with the null, suggesting that the alternative hypothesis is more likely to be true.

However, the p-value is not the probability of the null.
Using p-values for hypothesis testing

To obtain the p-value corresponding to the observed value of $\bar{x}$:
1. Compute the z-score of $\bar{x}$ as before, $z_{\text{obs}} = (\bar{x} - \mu_0) / (s / \sqrt{N})$.
2. Find the tail probability of $z_{\text{obs}}$ (probability of observing a value farther from $\mu_0$).
   a) If the alternative hypothesis is $\mu > \mu_0$: p-value = $\Pr(z > z_{\text{obs}})$.
   b) If the alternative hypothesis is $\mu < \mu_0$: p-value = $\Pr(z < z_{\text{obs}})$.
   c) If the alternative hypothesis is $\mu \neq \mu_0$: p-value = $\Pr(|z| > |z_{\text{obs}}|)$.

For a given significance level $\alpha$, we can reject the null when the p-value < $\alpha$. 
Using p-values for hypothesis testing

To obtain the p-value corresponding to the observed value of $\bar{x}$:
1. Compute the z-score of $\bar{x}$ as before, $z_{obs} = (\bar{x} - \mu_0) / (s / \sqrt{N})$.
2. Find the tail probability of $z_{obs}$ (probability of observing a value farther from $\mu_0$).
   a) If the alternative hypothesis is $\mu > \mu_0$: p-value = $Pr(z > z_{obs})$.
   b) If the alternative hypothesis is $\mu < \mu_0$: p-value = $Pr(z < z_{obs})$.
   c) If the alternative hypothesis is $\mu \neq \mu_0$: p-value = $Pr(|z| > |z_{obs}|)$.

Based on historical data, our team of programmers produces an average of 1000 lines of production-quality code per day. In the last 36 days, our team has used a new integrated development environment producing a mean of 1100 lines of production-quality code and standard deviation of 300 lines. Can we conclude that the new environment affects programmer productivity?

The z-score corresponding to $\bar{x} = 1100$ is $z = (1100 - 1000) / 50 = 2$.

\[
p-value = Pr(|z| > 2) = 1 - 2*F(2) = 0.0456.
\]

If $\alpha = 0.05$, we would reject the null since p-value < $\alpha$.
If $\alpha = 0.01$, we would not reject the null.
Using p-values for hypothesis testing

To obtain the p-value corresponding to the observed value of $\overline{x}$:
1. Compute the z-score of $\overline{x}$ as before, $z_{\text{obs}} = (\overline{x} - \mu_0) / (s / \sqrt{N})$.
2. Find the tail probability of $z_{\text{obs}}$ (probability of observing a value farther from $\mu_0$).
   a) If the alternative hypothesis is $\mu > \mu_0$: p-value = $\Pr(z > z_{\text{obs}})$.
   b) If the alternative hypothesis is $\mu < \mu_0$: p-value = $\Pr(z < z_{\text{obs}})$.
   c) If the alternative hypothesis is $\mu \neq \mu_0$: p-value = $\Pr(|z| > |z_{\text{obs}}|)$.

A computer supplies retail chain has a policy of only opening stores in communities where households spend more than $40 per year on computing supplies and equipment. A survey of 100 households in Monroeville finds that average expenditures in the sample are $40.50 with a standard deviation of $10. Is this strong evidence that the community spends more than $40?

The z-score corresponding to $\overline{x} = 40.50$ is $z = (40.50 - 40) / 1 = 0.5$.

$$p\text{-value} = \Pr(z > 0.5) = 0.5 - F(0.5) = 0.3085.$$ If $\alpha = 0.05$, we would not reject the null since $p\text{-value} \geq \alpha$. 
Comparing two populations

We can also make inferences comparing some parameter of two different populations, such as the population mean $\mu$ or the population proportion $p$.

Let us assume that we have a random sample from each population, and that these samples are drawn independently.

The average hourly wage of a random sample of 196 working women in Allegheny County is $8.21, with a standard deviation of $6.66. For a random sample of 204 working men, the average hourly wage is $12.96, with a standard deviation of $11.41. Is this sample evidence sufficient to conclude that there is a difference between the wages of men and women?

Let $\mu_1 =$ average hourly wage of men in Allegheny County. Let $\mu_2 =$ average hourly wage of women in Allegheny County.

We want to find confidence intervals for $\mu_1 - \mu_2$, and to test whether $\mu_1 - \mu_2 = 0$. 
Large-sample confidence intervals for the difference in means, $\mu_1 - \mu_2$

There is a probability of $c$ that $\mu_1 - \mu_2$ lies within $(\bar{x}_1 - \bar{x}_2) +/− z_c \sigma$.

<table>
<thead>
<tr>
<th>True difference</th>
<th>Observed difference</th>
<th>Number of std. dev. from mean</th>
<th>Std. dev. of observed difference</th>
</tr>
</thead>
</table>

How to obtain $\sigma$, the standard deviation of the observed difference?

**Answer:** the variance of $\bar{x}_1 - \bar{x}_2$ is the sum of the variance of $\bar{x}_1$ and the variance of $\bar{x}_2$.

$$\sigma^2 = (\sigma_1 / \sqrt{n_1})^2 + (\sigma_2 / \sqrt{n_2})^2 = (\sigma_1^2 / n_1) + (\sigma_2^2 / n_2) \approx (s_1^2 / n_1) + (s_2^2 / n_2).$$

$$\sigma = \sqrt{(s_1^2 / n_1) + (s_2^2 / n_2)}$$

There is a probability of $c$ that $\mu_1 - \mu_2$ lies within $(\bar{x}_1 - \bar{x}_2) +/− z_c \sqrt{(s_1^2 / n_1) + (s_2^2 / n_2)}$. 
Large-sample confidence intervals for the difference in means, $\mu_1 - \mu_2$

The average hourly wage of a random sample of 196 working women in Allegheny County is $8.21, with a standard deviation of $6.66. For a random sample of 204 working men, the average hourly wage is $12.96, with a standard deviation of $11.41. Is this sample evidence sufficient to conclude that there is a difference between the wages of men and women?

Let $\mu_1 = \text{average hourly wage of men in Allegheny County.}$
Let $\mu_2 = \text{average hourly wage of women in Allegheny County.}$

\[
\bar{x}_1 = 12.96 \quad \bar{x}_2 = 8.21 \\
s_1 = 11.41 \quad s_2 = 6.66 \\
n_1 = 204 \quad n_2 = 196
\]

There is a 95% probability that $\mu_1 - \mu_2$ lies within:

\[
(12.96 - 8.21) \pm 1.96 \sqrt{\frac{(11.41^2)}{204} + \frac{(6.66^2)}{196}}. \\
95\% \ CI = 4.75 \pm 1.96(0.930) = \[2.93, 6.57].
\]

There is a probability of $c$ that $\mu_1 - \mu_2$ lies within $(\bar{x}_1 - \bar{x}_2) \pm z_c \sqrt{\frac{(s_1^2)}{n_1} + \frac{(s_2^2)}{n_2}}.$
Large-sample hypothesis tests for the difference in means, $\mu_1 - \mu_2$

Do the means differ significantly? We want to test the alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ against the null hypothesis $H_0 : \mu_1 = \mu_2$.

If the null hypothesis was true, the observed difference $\bar{x}_1 - \bar{x}_2$ would follow a normal distribution with mean 0 and standard deviation $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

Z-score of the observed difference: $z_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$.

P-value for a two-sided test: $\Pr(|z| > |z_{\text{obs}}|) = 1 - 2*F(z_{\text{obs}})$.

Reject the null if p-value < $\alpha$.

\[ \bar{x}_1 = 12.96 \quad \bar{x}_2 = 8.21 \]
\[ s_1 = 11.41 \quad s_2 = 6.66 \]
\[ n_1 = 204 \quad n_2 = 196 \]

\[ \bar{x}_1 - \bar{x}_2 = 4.75 \quad \text{std. dev.} = 0.930 \]

Z-score = $4.75 / 0.930 = 5.11$

P-value $\approx 0.000$

We can reject the null hypothesis, and conclude that $\mu_1 \neq \mu_2$. 
Small-sample inference for the difference in means, $\mu_1 - \mu_2$

As in the one-population case, small-sample inference is more difficult because the Central Limit Theorem does not guarantee that the sample means $\bar{x}_1$ and $\bar{x}_2$ are normally distributed, and because $s_1$ and $s_2$ may not be accurate estimates of $\sigma_1$ and $\sigma_2$.

In order to do small-sample inference, we must make several simplifying assumptions: both populations must be approximately normally distributed, and must have equal variances $\sigma_1^2 = \sigma_2^2$. 

\[ \mu_1 \quad \mu_2 \]
Small-sample inference for the difference in means, $\mu_1 - \mu_2$

As in the one-population case, small-sample inference is more difficult because the Central Limit Theorem does not guarantee that the sample means $\bar{x}_1$ and $\bar{x}_2$ are normally distributed, and because $s_1$ and $s_2$ may not be accurate estimates of $\sigma_1$ and $\sigma_2$.

In order to do small-sample inference, we must make several simplifying assumptions: both populations must be approximately normally distributed, and must have equal variances $\sigma_1^2 = \sigma_2^2$.

For confidence intervals:

There is a probability of c that $\mu_1 - \mu_2$ lies in $(\bar{x}_1 - \bar{x}_2) \pm t_c \sigma$.

For hypothesis tests:

$t$-score of the observed difference: $t_{obs} = (\bar{x}_1 - \bar{x}_2) / \sigma$.

How to obtain $\sigma$, the standard deviation of the observed difference? How to obtain the number of degrees of freedom for the $t$-distribution?
Small-sample inference for the difference in means, $\mu_1 - \mu_2$

Solution: recall that we are assuming equal population variances, $\sigma_1^2 = \sigma_2^2 = \sigma_p^2$.

We must first estimate the pooled variance using the sample variances $s_1^2$ and $s_2^2$.

Pooled sample variance: $s_p^2 = \frac{(n_1 - 1)(s_1^2) + (n_2 - 1)(s_2^2)}{n_1 + n_2 - 2}$

$s_p^2$ is a weighted average of the sample variances $s_1^2$ and $s_2^2$, each weighted by its number of degrees of freedom $n_i - 1$.

Standard deviation of the observed difference:

$\sigma = \sqrt{s_p^2 / n_1} + (s_p^2 / n_2) = s_p \sqrt{1 / n_1} + (1 / n_2)$.

Total number of degrees of freedom: $n_1 + n_2 - 2$.

How to obtain $\sigma$, the standard deviation of the observed difference?

How to obtain the number of degrees of freedom for the t-distribution?
Small-sample inference for the difference in means, $\mu_1 - \mu_2$

The average hourly wage of a random sample of 16 working women in Allegheny County is $8.21, with a standard deviation of $6.66. For a random sample of 14 working men, the average hourly wage is $12.96, with a standard deviation of $11.41. Is this sample evidence sufficient to conclude that there is a difference between the wages of men and women?

Assuming equal variances, we compute the pooled sample variance as

$$s_p^2 = \frac{(15*6.66^2 + 13*11.41^2)}{28} = 84.2,$$

so $s_p = 9.18$. Then the standard deviation of the observed difference is

$$\sigma = 9.18 \sqrt{\left(\frac{1}{16}\right) + \left(\frac{1}{14}\right)} = 3.36.$$

Total number of degrees of freedom: $16 + 14 - 2 = 28$.

**Standard deviation of the observed difference:**

$$\sigma = \sqrt{s_p^2 / n_1} + s_p \sqrt{(1/n_1) + (1/n_2)}.$$

**Pooled sample variance:**

$$s_p^2 = \frac{((n_1 - 1)(s_1^2) + (n_2 - 1)(s_2^2))}{(n_1 + n_2 - 2)}$$
Small-sample confidence intervals for the difference in means, $\mu_1 - \mu_2$

The average hourly wage of a random sample of 16 working women in Allegheny County is $8.21, with a standard deviation of $6.66. For a random sample of 14 working men, the average hourly wage is $12.96, with a standard deviation of $11.41. Is this sample evidence sufficient to conclude that there is a difference between the wages of men and women?

Assuming equal variances, we compute the pooled sample variance as $s_p^2 = (15*6.66^2 + 13*11.41^2) / 28 = 84.2$, so $s_p = 9.18$. Then the standard deviation of the observed difference is $\sigma = 9.18 \sqrt{(1/16) + (1/14)} = 3.36$.

Total number of degrees of freedom: $16 + 14 - 2 = 28$.

For confidence intervals:

95% CI: $4.75 +/-(2.048)(3.36)$
= $[-2.13, +11.63]

99% CI: $4.75 +/-(2.763)(3.36)$
= $[-4.53, +14.03]$
Small-sample hypothesis tests for the difference in means, $\mu_1 - \mu_2$

The average hourly wage of a random sample of 16 working women in Allegheny County is $8.21, with a standard deviation of $6.66. For a random sample of 14 working men, the average hourly wage is $12.96, with a standard deviation of $11.41. Is this sample evidence sufficient to conclude that there is a difference between the wages of men and women?

Assuming equal variances, we compute the pooled sample variance as $s_p^2 = (15*6.66^2 + 13*11.41^2) / 28 = 84.2$, so $s_p = 9.18$. Then the standard deviation of the observed difference is $\sigma = 9.18 \sqrt{(1 / 16) + (1 / 14)} = 3.36$.

Total number of degrees of freedom: $16 + 14 - 2 = 28$.

For hypothesis tests:

$t$-score of the observed difference: $t_{obs} = (\bar{x}_1 - \bar{x}_2) / \sigma$.

$t$-score = $4.75 / 3.36 = 1.41$

$t_c$ for two-sided test ($\alpha = 0.05$, 28 dof): 2.048

Cannot reject $H_0$. 

Small-sample inference for the difference in means, $\mu_1 - \mu_2$

What if we do not believe that the population variances are equal?

It turns out that we can still do approximate inference, as the difference is still approximately t-distributed. The tricky part, though, is estimating the number of degrees of freedom.

The standard deviation of the observed difference is approximately the same as in the large-sample case, $\sigma = \sqrt{(s_1^2 / n_1) + (s_2^2 / n_2)}$.

However, the estimated number of degrees of freedom for the t-distribution is now:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2 / n_1)^2}{n_1 - 1} + \frac{(s_2^2 / n_2)^2}{n_2 - 1}}$$

In the case of equal sample sizes, $n_1 = n_2 = n$, we can use a much simpler approximation:

$$\nu = 2(n - 1)$$

This slide is optional material and will not be tested.
Large-sample confidence intervals for the difference in proportions

There is a probability of c that \( p_1 - p_2 \) lies within \((x_1 - x_2) +/- z_c \sigma\).

<table>
<thead>
<tr>
<th>True difference</th>
<th>Observed difference</th>
<th>Number of std. dev. from mean</th>
<th>Std. dev. of observed difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 - x_2 )</td>
<td>( x_1 - x_2 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How to obtain \( \sigma \), the standard deviation of the observed difference?

Answer: the variance of \( x_1 - x_2 \) is the sum of the variance of \( x_1 \) and the variance of \( x_2 \).

\[
\sigma^2 = \left(\sqrt{p_1(1 - p_1)/n_1}\right)^2 + \left(\sqrt{p_2(1 - p_2)/n_2}\right)^2 = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2} \approx \frac{(x_1(1 - x_1)/n_1) + (x_2(1 - x_2)/n_2)}

\[
\sigma = \sqrt{\frac{(x_1(1 - x_1)/n_1) + (x_2(1 - x_2)/n_2)}{2}}.

There is a probability of c that \( p_1 - p_2 \) lies within \((x_1 - x_2) +/- z_c \sqrt{\frac{(x_1(1 - x_1)/n_1) + (x_2(1 - x_2)/n_2)}{2}}\).
In a poll of 500 customers conducted two weeks after the implementation of a new computerized account system, you find that 49% are satisfied with the system. In a poll one month later, surveying an independent sample of 400 customers, you find the percentage of satisfied customers has increased to 53%. Can you conclude that support for the new system has increased?

Let \( p_1 \) = current proportion of customers who support the new system.
Let \( p_2 \) = original proportion of customers who support the new system.

\[
\bar{x}_1 = 0.53 \\
\bar{x}_2 = 0.49
\]

\[
\bar{x}_1 - \bar{x}_2 = 0.040, \quad \sqrt{\frac{(\bar{x}_1(1 - \bar{x}_1) / n_1) + (\bar{x}_2(1 - \bar{x}_2) / n_2)}{2}} = 0.034
\]

There is a 95% chance that \( p_1 - p_2 \) is between -0.026 and +0.106.

\[
0.04 +/– 1.96(0.034)
\]

There is a probability of \( c \) that \( p_1 - p_2 \) lies within

\[
(\bar{x}_1 - \bar{x}_2) +/– z_c \sqrt{\frac{(\bar{x}_1(1 - \bar{x}_1) / n_1) + (\bar{x}_2(1 - \bar{x}_2) / n_2)}{2}}.
\]
Large-sample hypothesis tests for the difference in proportions

If the null hypothesis was true with $p_1 = p_2$, then $\bar{x}_1 - \bar{x}_2$ would be normally distributed with mean 0 and standard deviation $\sigma$. How to find $\sigma$?

For proportions, $\sigma$ is different for confidence intervals and hypothesis tests:

1-population CI: $\sigma = \sqrt{\bar{x}(1 - \bar{x}) / N}$, 1-population HT: $\sigma = \sqrt{p_0(1 - p_0) / N}$

2-population CI: $\sigma = \sqrt{(\bar{x}_1(1 - \bar{x}_1) / n_1) + (\bar{x}_2(1 - \bar{x}_2) / n_2)}$.

2-population HT: $\sigma = \sqrt{p_0(1 - p_0)(1 / n_1 + 1 / n_2)}$ ← What is $p_0$?
Assume $p_1 = p_2 = p_0$.

The best estimate of $p_0$ under the null hypothesis is a weighted average of $\bar{x}_1$ and $\bar{x}_2$.

$$p_0 \approx (n_1 \bar{x}_1 + n_2 \bar{x}_2) / (n_1 + n_2)$$
Large-sample hypothesis tests for the difference in proportions

Has the proportion increased significantly? We want to test the alternative hypothesis $H_1: p_1 > p_2$ against the null hypothesis $H_0: p_1 \leq p_2$.

If the null hypothesis was true, the observed difference $\bar{x}_1 - \bar{x}_2$ would follow a normal distribution with mean 0 and standard deviation $\sqrt{p_0(1 - p_0)(1/n_1 + 1/n_2)}$.

z-score of the observed difference: $z_{obs} = (\bar{x}_1 - \bar{x}_2) / \sqrt{p_0(1 - p_0)(1/n_1 + 1/n_2)}$.

p-value for a one-sided test: $\Pr(z > z_{obs}) = 0.5 - F(z_{obs})$.

Reject the null if p-value < $\alpha$.

$x_1 = 0.53$
$n_1 = 400$

$x_2 = 0.49$
$n_2 = 500$

$\bar{x}_1 - \bar{x}_2 = 0.04$
$p_0 = 0.508$

std. dev. = 0.034

z-score = 0.04 / 0.034 = 1.19

p-value $\approx 0.1170$

We do not have sufficient evidence to reject the null hypothesis.
Paired differences

<table>
<thead>
<tr>
<th>Month</th>
<th>This yr</th>
<th>Last yr</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan.</td>
<td>35062</td>
<td>33956</td>
<td>1106</td>
</tr>
<tr>
<td>Feb.</td>
<td>27908</td>
<td>26544</td>
<td>1364</td>
</tr>
<tr>
<td>Mar.</td>
<td>18003</td>
<td>17443</td>
<td>560</td>
</tr>
<tr>
<td>Apr.</td>
<td>12544</td>
<td>12452</td>
<td>92</td>
</tr>
<tr>
<td>May</td>
<td>10708</td>
<td>9323</td>
<td>1385</td>
</tr>
<tr>
<td>Jun.</td>
<td>8322</td>
<td>7615</td>
<td>707</td>
</tr>
<tr>
<td>Jul.</td>
<td>8413</td>
<td>8222</td>
<td>191</td>
</tr>
<tr>
<td>Aug.</td>
<td>7857</td>
<td>8012</td>
<td>-155</td>
</tr>
<tr>
<td>Sep.</td>
<td>10190</td>
<td>9554</td>
<td>636</td>
</tr>
<tr>
<td>Oct.</td>
<td>15760</td>
<td>14220</td>
<td>1540</td>
</tr>
<tr>
<td>Nov.</td>
<td>28776</td>
<td>27639</td>
<td>1137</td>
</tr>
<tr>
<td>Dec.</td>
<td>43749</td>
<td>41500</td>
<td>2249</td>
</tr>
</tbody>
</table>

An online retailer of ski equipment wishes to compare monthly sales for the current year to last year’s monthly sales.

Mean sales for current year: $\bar{x}_1 = 18941$
Mean sales for last year: $\bar{x}_2 = 18040$
Std. dev. for current year: $s_1 \approx 12042$
Std. dev. for last year: $s_2 \approx 11552$
Number of samples: $n_1 = n_2 = 12$

95% CI for $\mu_1 - \mu_2 = 901 +/- 9990$
t-score = $901 / 4818 = .187$
p-value = 0.853, cannot reject $H_0$.

But we did better than last year, every month but August! What’s wrong here?
Paired differences

<table>
<thead>
<tr>
<th>Month</th>
<th>This yr</th>
<th>Last yr</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan.</td>
<td>35062</td>
<td>33956</td>
<td>1106</td>
</tr>
<tr>
<td>Feb.</td>
<td>27908</td>
<td>26544</td>
<td>1364</td>
</tr>
<tr>
<td>Mar.</td>
<td>18003</td>
<td>17443</td>
<td>560</td>
</tr>
<tr>
<td>Apr.</td>
<td>12544</td>
<td>12452</td>
<td>92</td>
</tr>
<tr>
<td>May</td>
<td>10708</td>
<td>9323</td>
<td>1385</td>
</tr>
<tr>
<td>Jun.</td>
<td>8322</td>
<td>7615</td>
<td>707</td>
</tr>
<tr>
<td>Jul.</td>
<td>8413</td>
<td>8222</td>
<td>191</td>
</tr>
<tr>
<td>Aug.</td>
<td>7857</td>
<td>8012</td>
<td>-155</td>
</tr>
<tr>
<td>Sep.</td>
<td>10190</td>
<td>9554</td>
<td>636</td>
</tr>
<tr>
<td>Oct.</td>
<td>15760</td>
<td>14220</td>
<td>1540</td>
</tr>
<tr>
<td>Nov.</td>
<td>28776</td>
<td>27639</td>
<td>1137</td>
</tr>
<tr>
<td>Dec.</td>
<td>43749</td>
<td>41500</td>
<td>2249</td>
</tr>
</tbody>
</table>

The assumption of independent samples is invalid. The counts for a given month are highly correlated between this year and last (high in winter, low in summer).

Notice that the variation from month to month is very large, compared to the relatively small difference between sample means.

We can reduce the variance of our sample by performing inference on the differences between this year’s counts and last year’s counts.

This method is only valid when we have matched pairs of datapoints; it cannot be used for independent samples.
Paired differences

<table>
<thead>
<tr>
<th>Month</th>
<th>This yr</th>
<th>Last yr</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan.</td>
<td>35062</td>
<td>33956</td>
<td>1106</td>
</tr>
<tr>
<td>Feb.</td>
<td>27908</td>
<td>26544</td>
<td>1364</td>
</tr>
<tr>
<td>Mar.</td>
<td>18003</td>
<td>17443</td>
<td>560</td>
</tr>
<tr>
<td>Apr.</td>
<td>12544</td>
<td>12452</td>
<td>92</td>
</tr>
<tr>
<td>May</td>
<td>10708</td>
<td>9323</td>
<td>1385</td>
</tr>
<tr>
<td>Jun.</td>
<td>8322</td>
<td>7615</td>
<td>707</td>
</tr>
<tr>
<td>Jul.</td>
<td>8413</td>
<td>8222</td>
<td>191</td>
</tr>
<tr>
<td>Aug.</td>
<td>7857</td>
<td>8012</td>
<td>-155</td>
</tr>
<tr>
<td>Sep.</td>
<td>10190</td>
<td>9554</td>
<td>636</td>
</tr>
<tr>
<td>Oct.</td>
<td>15760</td>
<td>14220</td>
<td>1540</td>
</tr>
<tr>
<td>Nov.</td>
<td>28776</td>
<td>27639</td>
<td>1137</td>
</tr>
<tr>
<td>Dec.</td>
<td>43749</td>
<td>41500</td>
<td>2249</td>
</tr>
</tbody>
</table>

Mean of differences: $\bar{x}_d = 901$
Std. dev. of differences: $s_d = 692$
Number of differences: $n_d = 12$

We can now do one-population inference for the population of differences.

Assumption: sample differences are sampled at random from the target population of differences.

For small samples, we must also assume that the population of differences is approximately normally distributed.
### Paired differences

<table>
<thead>
<tr>
<th>Month</th>
<th>This yr</th>
<th>Last yr</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan.</td>
<td>35062</td>
<td>33956</td>
<td>1106</td>
</tr>
<tr>
<td>Feb.</td>
<td>27908</td>
<td>26544</td>
<td>1364</td>
</tr>
<tr>
<td>Mar.</td>
<td>18003</td>
<td>17443</td>
<td>560</td>
</tr>
<tr>
<td>Apr.</td>
<td>12544</td>
<td>12452</td>
<td>92</td>
</tr>
<tr>
<td>May</td>
<td>10708</td>
<td>9323</td>
<td>1385</td>
</tr>
<tr>
<td>Jun.</td>
<td>8322</td>
<td>7615</td>
<td>707</td>
</tr>
<tr>
<td>Jul.</td>
<td>8413</td>
<td>8222</td>
<td>191</td>
</tr>
<tr>
<td>Aug.</td>
<td>7857</td>
<td>8012</td>
<td>-155</td>
</tr>
<tr>
<td>Sep.</td>
<td>10190</td>
<td>9554</td>
<td>636</td>
</tr>
<tr>
<td>Oct.</td>
<td>15760</td>
<td>14220</td>
<td>1540</td>
</tr>
<tr>
<td>Nov.</td>
<td>28776</td>
<td>27639</td>
<td>1137</td>
</tr>
<tr>
<td>Dec.</td>
<td>43749</td>
<td>41500</td>
<td>2249</td>
</tr>
</tbody>
</table>

Mean of differences: $\bar{x}_d = 901$

Std. dev. of differences: $s_d = 692$

Number of differences: $n_d = 12$

Confidence interval for $\mu_d = \bar{x}_d +/– t_c \left( \frac{s_d}{\sqrt{n_d}} \right)$

$n_d – 1$ degrees of freedom

95% CI = $901 +/– 2.201(199.8) = [461, 1341]$

Testing $\mu_d \neq 0$ (same as $\mu_1 – \mu_2 \neq 0$)

$t$-score = $\frac{\bar{x}_d}{\left( \frac{s_d}{\sqrt{n_d}} \right)} = \frac{901}{199.8} = 4.51$

$p$-value $\approx .000$

We can reject the null hypothesis and conclude that $\mu_1 \neq \mu_2$. 
### When to use paired differences?

<table>
<thead>
<tr>
<th>Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparing daily sales for two restaurants for the same set of 30 days.</td>
</tr>
<tr>
<td>Comparing salaries of male and female movie stars, sampling 50 of each.</td>
</tr>
<tr>
<td>Comparing the average reaction time of 50 subjects dosed with caffeine and 50 patients without caffeine.</td>
</tr>
<tr>
<td>Comparing the average stress on a car’s front and back wheels.</td>
</tr>
<tr>
<td>Comparing daily sales for two restaurants, choosing an independent set of 30 days for each restaurant.</td>
</tr>
<tr>
<td>Comparing salaries of male and female movie stars, matching each actor to an actress with similar experience, fame, etc.</td>
</tr>
<tr>
<td>Comparing each subject’s reaction time with and without caffeine.</td>
</tr>
<tr>
<td>Comparing the number of attempted network intrusions before and after installing a new firewall.</td>
</tr>
</tbody>
</table>