SUPPOSE

we think the mean value of a population of SAT scores is \( \mu = 455 \)

we can take a sample of the population and calculate the sample mean of SAT scores \( \bar{X} = 535 \)

we can make some statement about how rare it is to get \( \bar{X} = 535 \)

(what we did last time)

\textbf{or}

we can make a statement about how unreasonable it is that our original thought is true!

HYPOTHESIS TESTING

in hypothesis testing we consider how reasonable a hypothesis is, given the data that we have

if the hypothesis is reasonable (consistent with the data), we assume it could be true

if the hypothesis is unreasonable (inconsistent with the data), we assume it is false

deciding on what hypotheses to test is critically important!

HYPOTHESIS

conjecture about one or more population parameters

\[ \text{e.g.} \]

\( \mu = 455 \)

\( \sigma = 3.5 \)

\( r = 0.76 \)

\[
\text{in inferential statistics we always test the null hypothesis } H_0
\]

NULL HYPOTHESIS

\( H_0 \) is the assumption of no relationship, or no difference

\[ \text{e.g.} \]

\( H_0 \): no relationship between variables

\( H_0 \): no difference between treatment groups

the alternative hypothesis, \( H_a \) is the other possibility

\[ \text{e.g.} \]

\( H_a: \mu \neq 455 \)

\( H_0: \mu = 455 \)

(\text{does not say what } \mu \text{ is, but says what it is not!})
**NULL HYPOTHESIS**

what’s wrong with herbal medicines?
nothing necessarily, but I don’t know that they are any good (and they may be bad)
lots of reports that they help people (but how can they be sure)
need to start by assuming that a medicine does nothing, and prove that the assumption is false!
anecdotal reports are just about worthless

**STATE THE HYPOTHESIS**

before doing anything else, we need to make certain that we understand the tested hypothesis
for the SAT example

\[ H_0 : \mu = 455 \]

\[ H_a : \mu \neq 455 \]
sometimes this is the most difficult step in designing an experiment
to start, we will worry only about hypotheses about the population mean, \( \mu \)

**SIGNAL DETECTION**

The task is almost the same as deciding whether a measurement came from a noise-alone (null hypothesis) distribution or a signal-and-noise (alternative hypothesis) distribution
How well you can do is determined by the signal-to-noise ratio (\( d' \)), but that value is typically unknown
we set a criterion using only the null hypothesis (noise-alone distribution)

**CRITERION**

we will examine the data to see if we should reject \( H_0 \)
we will do that by comparing the sample mean, \( \bar{X} \), to the hypothesized value of the population mean, \( \mu \)
the bottom-line is whether \( \bar{X} \) is sufficiently different from \( \mu \) to reject \( H_0 \)
but we have to consider four things to quantify the term sufficiently different

- standard scores
- errors in hypothesis testing
- level of significance
- region of rejection

**STANDARD SCORES**

we previously used standard scores to indicate how much a given scores deviates from a distribution mean
We do the same kind of thing here, but we want to know how a sample mean, \( \bar{X} \) deviates from what the sampling distribution would be if the null hypothesis is true
We give the standard score a special term:

\[ t = \frac{\bar{X} - \mu}{s_{\bar{X}}} \]

We compute everything else using the sampling distribution of this \( t \) value: the \( t \) distribution, which is similar to a normal distribution with fatter tails and requires degrees of freedom:

\[ df = n - 1 \]
DECISIONS

after deciding to reject or not reject $H_0$ there are four possible situations

- A true null hypothesis is rejected. (False alarm)
- ** A true null hypothesis is not rejected. (Correct rejection)
- A false null hypothesis is not rejected. (Miss)
- ** A false null hypothesis is rejected. (Hit)

errors are unavoidable

we want to minimize the probability of making errors, given the particular data set we have

ERRORS

two types of errors:

- ** Type I error**: when we reject a true null hypothesis.
- ** Type II error**: when we do not reject a false null hypothesis.

<table>
<thead>
<tr>
<th>Decision made</th>
<th>State of nature</th>
<th>$H_0$</th>
<th>$H_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>True</td>
<td>$H_0$</td>
<td>$H_a$</td>
</tr>
<tr>
<td>Do not reject</td>
<td>False</td>
<td>$H_0$</td>
<td>$H_a$</td>
</tr>
</tbody>
</table>

generally, decreasing the probability of making one type of error increases the probability of making the other type of error

ERRORS

suppose you have a new, untested, and expensive treatment for cancer

you run a test to judge whether the drug is better than existing drugs

if you reject $H_0$, indicating that the drug is more effective, when in fact it is not, people will spend a lot of money for no reason (Type I error)

if you fail to reject $H_0$, indicating that the drug is not effective, when in fact it is, people will not use the drug (Type II error)

scientific research tends to focus on avoiding Type I errors

SIGNIFICANCE LEVEL

alpha ($\alpha$) level

indicates probability of Type I error

frequently we choose $\alpha = 0.05$ or $\alpha = 0.01$

that is, the corresponding decision to reject $H_0$ may produce a Type I error 5% or 1% of the time

a statement about how much error we will accept

usually chosen before the data is gathered

depends upon use of the analysis

REGION OF REJECTION

$\alpha$ is a probability

it identifies how much risk of Type I error we are willing to take (rejecting $H_0$ when it is true)

consider our example of SAT scores

$H_0 : \mu = 455$

suppose we also know the sample standard deviation

$s = 100$

and our sample size is $n = 144$

we know that the sampling distribution of $t$ is:

1. A $t$ distribution with $df = n - 1 = 143$.
2. Has a mean of $\mu = 0$, if $H_0$ is true.
3. Has a standard error of the mean $s \bar{X} = \frac{s}{\sqrt{n}} = \frac{100}{\sqrt{144}} = 8.33$
REGION OF REJECTION

area under the curve represents the probability of getting the corresponding t values, if the \( H_0 \) is true.

the extreme tails of the sampling distribution correspond to what should be very rare t values, and thus very rare sample means.

TEST STATISTIC

if the t-score is beyond \( \pm 1.977 \), it is very unlikely to have occurred if the \( H_0 \) is true.

we have the following data:

- \( \mu = 455 \), \( H_0 \)
- \( n = 144 \), sample size
- \( \bar{X} = 535 \), observed value for sample statistic
- \( s = 100 \), value of the standard deviation of the population
- \( s_{\bar{X}} = 8.33 \), standard error (calculated earlier)

from this we can calculate the t-score

\[
t = \frac{\bar{X} - \mu}{s_{\bar{X}}}
\]

\[
t = \frac{535 - 455}{8.33} = 9.60
\]

the standard score is the test statistic for testing \( H_0 \) about a population mean

DECIDING ABOUT \( H_0 \)

compare the test statistic to the critical value

\[
9.60 > 1.977
\]

indicates that the sample mean \( \bar{X} \) is extremely rare, given the assumed population mean \( \mu \), by chance (random sampling)
another way to do it (advocated by your text) is to use the $t$-value to compute the probability of getting a $t$-value more extreme than what you found.

$p$-value

$t$ distribution calculator

We find $p \approx 0$

Since the probability is small ($< .05$), then we conclude that the $H_0$ is probably not true.

$p$-value

Since the $p$ value is smaller than the $\alpha$ we set, we reject $H_0 : \mu = 455$ in favor of the alternative hypothesis $H_a : \mu \neq 455$ but there is still a chance that $H_0$ is true!

Decisions

null hypothesis

rejecting $H_0$

Type I error

Type II error

Why clinical studies use thousands of subjects.