Chapter Seven

Hypothesis Testing with *Z* and *T*

Introduction to Hypothesis Testing

P Values

Critical Values

Within-Subjects Designs

Between-Subjects Designs

Hypothesis Testing

An **alternate hypothesis (H_1)** is a researcher's prediction. It states that the population parameter is different from a given value or from that of another population, and it usually states the direction of the difference.

A **null hypothesis (H_0)** is the complement of the alternate hypothesis.

The **tail** of a test is the direction in which the alternate hypothesis predicts the results will turn out.

The examples below refer to the question of how meditation affects concentration.

Hypothesis	Right-tailed	Left-Tailed	Two-tailed
Alternate	Meditation increases concentration.	Meditation decreases concentration.	Meditation affects concentration.
Null	Meditation does not increase concentration.	Meditation does not decrease concentration.	Meditation does not affect concentration.

Hypothesis testing is using sample data and statistical methods to try to reject the null hypothesis.

The null hypothesis is either rejected or not rejected. If it is not rejected, this does not mean that it is "accepted." As an analogy, consider the difference in meanings in the statements *We know there is not life on other planets* and *We do not have evidence of life on other planets*.

Type I and Type II Errors

If the data in a study strongly support the hypothesis, they are said to be **statistically significant**, and the null hypothesis is rejected. However, since sample data only provide estimates for population parameters, **hypothesis testing cannot "prove" anything**. There is always the possibility that a finding, or a lack of finding, is a conicidence in the sample and not representative of the population overall.

A **type I error** is making a false claim, that is, rejecting the null hypothesis when it is actually true.

A **type II error** is not making a claim when one should be made, that is, not rejecting the null when it is actually false.

Possibility	Reject H_0	Do not reject H_0
H_0 is actually true.	Type I error	correct
H_0 is actually false.	correct	Type II error

Rougly speaking, a type I error is finding something that isn't really there, and a type II error is missing something that is there, as in the example below of a clinical trial for a new drug therapy.

Possibility	Reject H_0	Do not reject H_0
The drug is not effective.	A worthless drug is produced.	A worthless drug is not produced.
The drug is effective.	A valuable drug is produced.	A valuable drug is not produced.

Power

The **power** of a test is the probability that a false null hypothesis will be rejected.

Three factors lead to large z or t scores and thus increased power.

Factor	Example of higher power	Example of lower power
Big difference	Boys read an average of 386 words per minute, and girls read an average of 463 words per minute.	Boys read an average of 386 words per minute, and girls read an average of 398 words per minute.
Large sample	80 boys and 74 girls	10 boys and 13 girls
Low variance	$s_1 = 54, s_2 = 48$	$s_1 = 144, s_2 = 148$

In addition, one-tailed tests are more powerful than two-tailed tests. This is because when a p value is calculated for a two-tailed test, it is twice as big.

P Values

The *P* value of a sample statistic is the probability that another sample of the same size would support the the prediction at least as well, given the null hypothesis is true.

Prediction	Outcome	*P* Value	Conclusion
A particular coin lands on heads more than tails.	6 out of 9 flips are heads.	$\binom{9}{6}\left(\frac{1}{2}\right)^9 + \binom{9}{7}\left(\frac{1}{2}\right)^9 + \binom{9}{8}\left(\frac{1}{2}\right)^9 + \binom{9}{9}\left(\frac{1}{2}\right)^9 \approx 25.4\%$	The prediction was correct, but not by enough to be convincing that the coin is weighted. 6 out of 9 heads could easily be a coincidence.
A particular coin lands on heads more than tails.	9 out of 9 flips are heads.	$\left(\frac{1}{2}\right)^9 \approx 0.2\%$	This coin seems to be weighted towards heads. 9 heads out of 9 could have been a coincidence, but this is not likely.
SAT math scores on average are higher than 500 ($\sigma = 100$).	The average of 250 random scores is 512.	$z = \frac{512 - 500}{100 / \sqrt{250}} = 1.90$ $P(z > 1.90) = 2.9\%$	The average SAT math score seems to be above 500. Although the sample mean was only slightly higher than 500, the sample was large, making the difference unlikely to be coincidental.

Using the *z* table and methods from previous chapters, it is easy to calculate the *p* value for a sample based on its *z* score. For means, $\boldsymbol{z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}}$. For proportions, $\boldsymbol{z = \frac{\hat{p} - p}{\sqrt{pq/n}}}$.

Common misunderstandings about *p* values

False understanding	Correct understanding	Non-statistics example
A *p* value is the probability that a type I error was made.	A *p* value is the probability of reaching this sample result, given the null is true. This is not the same as the probability that the null is true, given this sample result.	The probability that a criminal is a male is very different from the probability that a male is a criminal.
***P* can be calculated using the data that were used to form the hypothesis.**	*P* values only have meaning for events that were predicted. It's important not to make a "prediction" about an event that you know has already happened.	Players call their shots in pool. If a player makes a shot without calling it, it is assumed that it was a coincidence rather than skill, and it doesn't count.

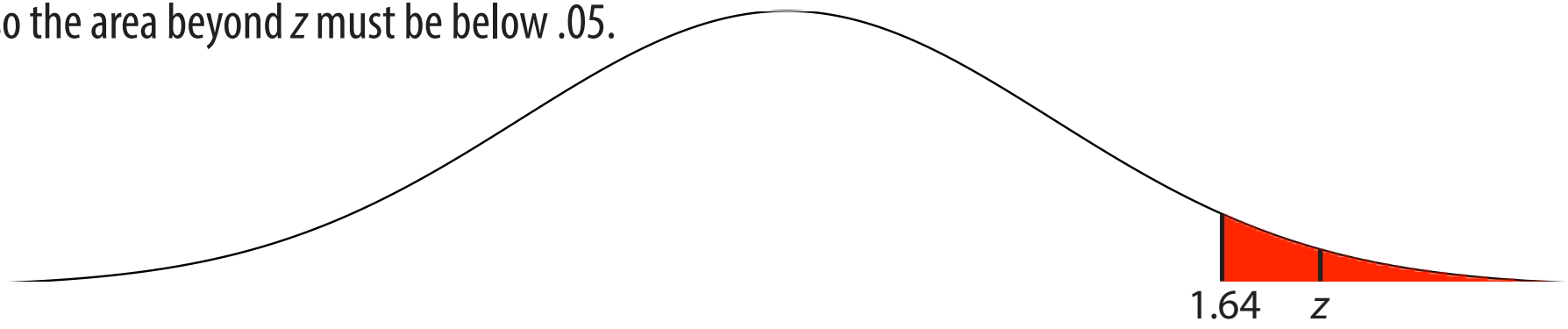
Critical Values and Statistical Significance

The **level of significance** α (alpha) is how low the *p* value needs to be for the data to be statistically significant, allowing the researcher to reject H_0. In scientific research, the globally accepted standard is $\alpha = .05$, although a different level can be used if it is preregestered before the research is conducted. Always use a 5% level of significance for this course.

If *p* is less than .05, the null hypothesis is rejected no matter what *p* is. Therefore, rather than calculating *p*, it is sufficient to simply identify the **critical value** at which *p* will be .05. For a *z* test, this value is $z_0 = 1.64$ for right-tailed or $z_0 = \pm 1.96$ for two-tailed. For a *t* test, the critical value can be looked up in the *t* table by degrees of freedom.

The region under the curve beyond the critical value is the **critical region** and has an area of .05. If the sample statistic has reached the critical value, then it is in the critical region, *p* is less than .05, the data are statistically significant, and the null hypothesis is rejected.

In the example below, the sample statistic *z* is in the critical region. The area of the critical region is .05, so the area beyond *z* must be below .05.



Rejecting the Null Hypothesis

Each statement below implies the one below it. Therefore, they essentially all mean the same thing.

Statement	Meaning
The sample statistic is greater* than the critical value.	*z* is greater* than 1.64, or *t* or another sample statistic is greater* than the critical value for that statistic for the given number of degrees of freedom. * or less, depending on the tails
The sample statistic is in the critical region.	The sample statistic (*z*, *t*, etc.) falls within the 5% of the curve that is shaded in the tail(s) past the critical value(s).
The *p* value is less than .05.	If the null hypothesis is true, there is less than a 5% chance that a new random sample of the same size would support the alternate hypothesis as strongly as the current one did.
The data are statistically significant.	The sample size is large enough, the results are different enough from each other, and there is low enough variance for the researchers to be convinced that their findings are not merely a coincidence in their sample.
The null hypothesis is rejected.	The researchers claim that their prediction, which turned out true in their sample, is valid for the population overall. However, they may be making a type I error.

Between-Subjects and Within-Subjects Designs

Data collected from a sample can be compared to data from another sample or to data collected again from the same sample under different conditions. Within-subjects designs are not always applicable, but they are more powerful than between-subjects designs and should be used when feasible.

Design	Description	Example	Advantage
Between-Subjects	A different group of participants is used for condition B than was used for condition A.	Some participants memorize a word list by reading it, and some participants memorize a word list by hearing it.	• Avoids sequence effects: Participants cannot be influenced by their earlier participation. • It may be impractical, immoral, or impossible to have each participant take part in each condition.
Within-Subjects	The same group of participants takes part in both conditions.	Every participant memorizes one word list by reading it and another word list by hearing it.	• More powerful: Since each participant is being compared against themself, an outlier in one condition will likely be an outlier in the other condition as well, making the difference not an outlier. This makes for lower variance, and thus a more powerful test.

Sequence Effects

Sequence effects are effects of one condition of a study on a later condition, such as those below.

Sequence Effect	Example of a dependent variable that could be affected
Fatigue	time to run a mile
Practice	free-throw percentage
Boredom	score on a timed math test
Interference	number of words remembered
Environmental Changes	driving speeds (if weather changes)
Knowledge about Procedure	whether or not the gorilla was noticed

When possible, sequence effects are mitigated by **counterbalancing**, in which some participants take part in condition A first and the others take part in condition B first. If the the participants are randomly assigned to which condition they will take part in first, then the within-subjects design is a true experiment (rather than a quasi-experiment), eliminating additional possible confounding variables.

Two-sample tests

Z tests and *t* tests can be done for one or two samples.

Between-subjects designs use two-sample tests. Within-subjects designs use a one-sample test even though there are two conditions, because only one mean (the mean difference) is being tested.

Parameter	One-sample test	Two-sample test
Mean	$t = \frac{\bar{x} - \mu}{s\sqrt{\frac{1}{n}}}$	$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $s = \sqrt{\frac{df_1(s_1^2) + df_2(s_2^2)}{df_1 + df_2}}$
Proportion	$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}}$

The two-sample *t*-test formula combines s_1 and s_2 into a **pooled** estimate of σ by taking the weighted average of the sample standard deviations based on degrees of freedom. If there is reason to assume that σ would be significantly different for one population than the other, then this formula is only valid for tests in which the two sample sizes are similar. (Other methods, not covered in this course, can be used for other situations.)

If σ is known, *z* is calculated instead of *t*.

Calculator tests of means

Z tests and *t* tests of one or two means can be done on the calculator.

On calculator	One-sample test*	Two-sample test
Test	`Z-Test...` if σ is known `T-Test...` if σ is unknown	`2-SampZTest` if both σ's are known `2-SampTTest` if not
Input	Enter data in `L1` and choose `Data`, or choose `Stats` and enter the sample mean, the population mean being tested (for H_0), the standard deviation, and the sample size.	Enter data in `L1` and `L2` and choose `Data`, or choose `Stats` and enter the sample means, the standard deviations, and the sample sizes.
Tails setting	$\mu>\mu_0$ for right-tailed $\mu<\mu_0$ for left-tailed $\mu\neq\mu_0$ for two-tailed	$\mu_1>\mu_2$ for right-tailed *(testing if the first population mean is higher than the second population mean)* $\mu_1<\mu_2$ for left-tailed *(testing if the first population mean is lower than the second population mean)* $\mu_1\neq\mu_2$ for two-tailed

* including within-subjects designs

Calculator tests of proportions

Z tests of one or two proportions can be done on the calculator.

On calculator	One-sample test	Two-sample test
Test	`1-PropZTest...`	`2-PropZTest...`
Input	Enter `p0` for the population proportion being tested, `n` for the sample size, and `x` for the number of successes in the sample.	Enter `n1` for the size of the first sample, and `x1` for the number of successes in that sample. Then do the same for `n2` and `x2` for the second sample.
Tails setting	`prop>p0` for right-tailed `prop<p0` for left-tailed `prop≠p0` for two-tailed	`p1>p2` for right-tailed *(testing if the first population proportion is higher than the second)* `p1<p2` for left-tailed *(testing if the first population proportion is lower than the second)* `p1≠p2` for two-tailed