TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI

# Statistics: Hypothesis Testing
# MATHematics MAGnification™

Dr. Ralph deLaubenfels

TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI TSI

Teacher-Scholar Institute

Columbus, Ohio

2019

## STATISTICS: HYPOTHESIS TESTING MAGNIFICATION

This is one of a series of very short books on math, statistics, and physics called "Math Magnifications." The "magnification" refers to focusing on a particular topic that is pivotal in or emblematic of mathematics.

## OUTLINE

Statistical inference comes primarily in two forms: confidence intervals and hypothesis testing. This magnification, after some general results and motivation, will talk about a special case of hypothesis testing; a prior magnification ([2]) talked about the same special case for confidence intervals. A future magnification will show how the same constructions work for most other popular confidence intervals and hypothesis tests.

This magnification will perform hypothesis testing for the mean of a normal population, with standard deviation known. As with confidence intervals ([2]), some of the desirabilities of large samples will be exhibited.

Prerequisites for this magnification are algebra ([5] is more than sufficient), the basic terminology of statistics, as in [4] or [3], and familiarity with normal populations, such as is covered in [2, Chapter 2]. Some more general knowledge of probability is helpful, but [2, Chapter 2], along with the appendix attached to this magnification, should be sufficient.

## 1. INTRODUCTION: NULL HYPOTHESIS and ALTERNATIVE HYPOTHESIS

Hypothesis testing is a formal statistical formulation for drawing conclusions or making decisions. Said formulation will make explicit and quantified reasoning that we do instinctively and perhaps unconsciously. It is better to have such a formulation, to bring our mental processes out in the open, if only to identify limitations and possible flaws. This set-up will place us on the interface of ideas and actions.

Throughout this magnification, the *probability* of an event $A$ is denoted $P(A)$.

**Definition 1.1.** A **hypothesis** is an assertion or statement of (alleged) fact.

**Examples 1.2.** Here are some examples of hypotheses.

(a.) You committed a crime.

(b.) You didn't commit a crime.

(c.) Eating fish makes people smarter.

(d.) Mars is about to explode.

(e.) A coin is unfair, meaning that, when you flip said coin, the probability of the coin coming up heads does not equal the probability of the coin coming up tails.

**Definitions 1.3.** A **hypothesis test** consists of a pair of mutually exclusive hypotheses, denoted $H_0$, the **null hypothesis**, and $H_a$, the **alternative hypothesis**.

Examples 1.2(a.) and (b.) are an example of a pair of mutually exclusive hypotheses.

**Example 1.4.** Here's a stressful example, representing a claim that you cheated on an exam.

$H_0$: You didn't cheat on an exam.

$H_a$: You cheated on an exam.

**1.5. Customary choices of null and alternative hypotheses.** The null hypothesis $H_0$ should be a *default*, something you believe, until decisively proven otherwise. The alternative hypothesis $H_a$ is believed grudgingly, only after strong evidence has appeared; addressing $H_a$ requires the collection of data, as in [3].

Thus Example 1.4 is typical, at least in a culture with sufficient civility to dislike frivolous, inflammatory accusations and false convictions. More generally, the legal concept of "presumption of innocence" or "innocent until proven guilty" implies that the hypothesis in Examples 1.2(b.) should be $H_0$, a null hypothesis, in any hypothesis test concerned with the possible commission of a crime.

Much more generally, if there is a provocative, unpleasant, or in some way problematic claim to be tested, said claim becomes the alternative hypothesis $H_a$. See Examples 1.6 and 1.7.

**1.6. More Examples.** Set up each of the following claims as hypothesis tests.

(a.) Canadian wolverines are fatter than American wolverines.

**Solution.**

$$H_0: \text{ Canadian wolverines are the same weight as American wolverines}$$

$$H_a: \text{ Canadian wolverines are heavier than American wolverines}$$

(b.) Sleeping on the floor will make you rich.

**Solution.**

$$H_0: \text{ people sleeping on the floor have the same annual average salary as all people}$$
$$H_a: \text{ people sleeping on the floor have a higher annual average salary than all people}$$

(c.) Colored lights will cure glaucoma.

**Solution.**

$$H_0: \text{ people exposed to colored light have the same chance of getting glaucoma as all people}$$
$$H_a: \text{ people exposed to colored light have a smaller chance of getting glaucoma than all people}$$

It is preferable, where possible, to describe everything with specific parameters, to set up future quantifications. Here are some examples.

**Examples 1.7.** Set up each of the following as hypothesis tests.

(a.) I accuse you of using an unfair coin (see Examples 1.2(e.)).

**Solution.** We can and should imitate Example 1.4, since use of an unfair coin usually implies cheating:

$H_0$: The coin is fair.

$H_a$: The coin is not fair.

Here we could choose the parameter

$$p \equiv P(\text{heads}), \text{the probability of getting heads whenever the coin is flipped.}$$

Then the coin is fair if and only if $p = 0.5$, thus we may set up the following hypothesis test.

$$H_0 : p = 0.5 \quad H_a : p \neq 0.5.$$

As in 1.5, this hypothesis test places the burden of proof on me: $H_0$, meaning your use of a fair coin, is assumed, until I prove $H_a$, my serious accusation of cheating by using an unfair coin, conclusively.

(b.) "Meds 'R Us" (MRU) claims their medication, "Taller Than Thou" (TTT) makes people taller (on average). If we assume the average human height is 70 inches, test the claim made by MRU.

**Solution.** Let $\mu$ be the average height of people taking TTT. Here is our hypothesis test.

$$H_0 : \mu = 70 \quad H_a : \mu > 70.$$

Notice that MRU's claim became $H_a$. The claim needs strong evidence, before we spend all our money trying to create a basketball team with TTT. The default, $H_0$, asserts that TTT does nothing.

(c.) Same as (b.), except the claim is now that TTT makes people shorter (on average).

**Solution.** With $\mu$ again the average height of people taking TTT, here is our hypothesis test.

$$H_0 : \mu = 70 \quad H_a : \mu < 70.$$

(d.) Same as (b.), except the claim is now "TTT will change people's height (on average)."

**Solution.** Again $\mu$ is the average height of people taking TTT:

$$H_0 : \mu = 70 \quad H_a : \mu \neq 70.$$

**1.8. Desired conclusion of hypothesis test** in Definitions 1.3. You should conclude either

$$(1.)\ \text{Reject } H_0;$$

or

$$(2.)\ \text{Don't reject } H_0.$$

In more detail:

$$(1.)\ \text{Reject } H_0 \text{ in favor of } H_a;\ \text{there is compelling evidence to accept } H_a;$$

or

$$(2.)\ \text{Don't reject } H_0 \text{ in favor of } H_a;\ \text{there is insufficient evidence to accept } H_a.$$

.

This is sometimes called **testing $H_0$ against $H_a$**

For example, in Example 1.4 the possible conclusions are either "There is compelling evidence to conclude that you cheated on an exam," or "There is insufficient evidence to conclude that you cheated on an exam."

The latter conclusion is *not* guaranteeing that you didn't cheat; it only states that we need more evidence before all the unpleasantness of convicting. Much more generally, it is misleading to say we "accept $H_0$;" we can only fail to reject $H_0$, leaving us in the same state of ignorance as before the hypothesis test.

## 2. P-VALUES: HOW WEIRD IS MY DATA?

After setting up our hypothesis test as in 1.3 and 1.5, we collect some data. The most informative single number describing this data, relevant to testing $H_0$ against $H_a$, as in 1.8, is known as the *P-value* of the data.

**Definition 2.1.** For a given hypothesis test, the **P-value**, or **observed significance level**, of some data you've collected is the probability, assuming $H_0$, of getting data that seems to favor $H_a$ at least as much as the data you got.

Notice that our default of $H_0$ is implicit in the definition of P-value where probabilities are calculated under the assumption of $H_0$.

**Examples 2.2.** Get P-values for each of the following hypothesis tests and data.

(a.) I claim that a mysterious coin from the Dawn of Time is weighted so that, on each flip, we are more likely to get heads than tails.

I collect data by flipping said coin ten times; I get all heads.

**Solution.** Let $p \equiv P(\text{heads})$, the probability of getting heads when I flip the coin of mystery. Following 1.5, 1.6, and 1.7, especially similar to Examples 1.7(a.), my hypothesis test is

$$H_0 : p = 0.5 \quad H_a : p > 0.5.$$

Our P-value is

$$P(10 \text{ heads in ten flips of a fair coin}),$$

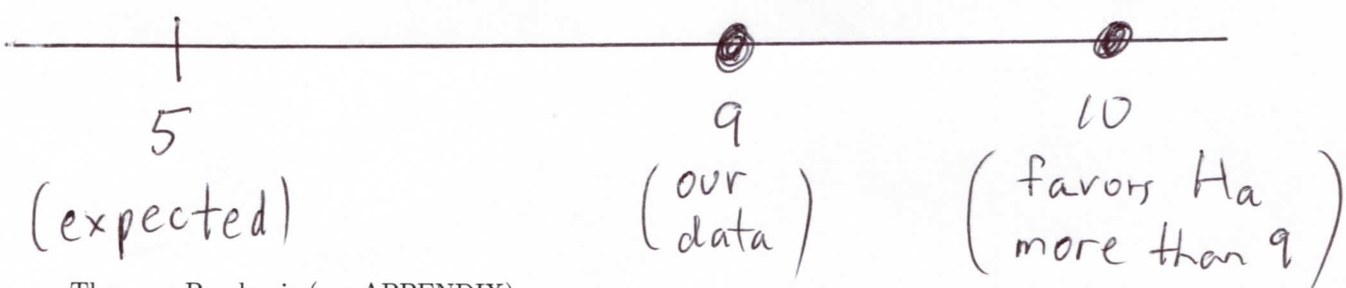
the probability of getting ten heads in ten flips of a fair coin, which equals (see APPENDIX) $\binom{10}{10}(\frac{1}{2})^{10} = (\frac{1}{2})^{10} \sim 9.77 \times 10^{-4} = 0.0977\%$.

(b.) Same as (a.), except I get nine heads in ten flips.

**Solution.** We have the same hypothesis test as in (a.). Our P-value includes

$$P(9 \text{ heads in ten flips of a fair coin}),$$

the probability of getting the data we got, but it is not limited to that, because getting 10 heads would cause us to favor $H_a$ even more than getting 9 heads, since, when flipping a fair coin ten times, we expect (on average) to get $0.5 \times 10 = 5$ heads. See the drawing directly below; note that 10 is at least as far away from 5 as 9.




Thus our P-value is (see APPENDIX)

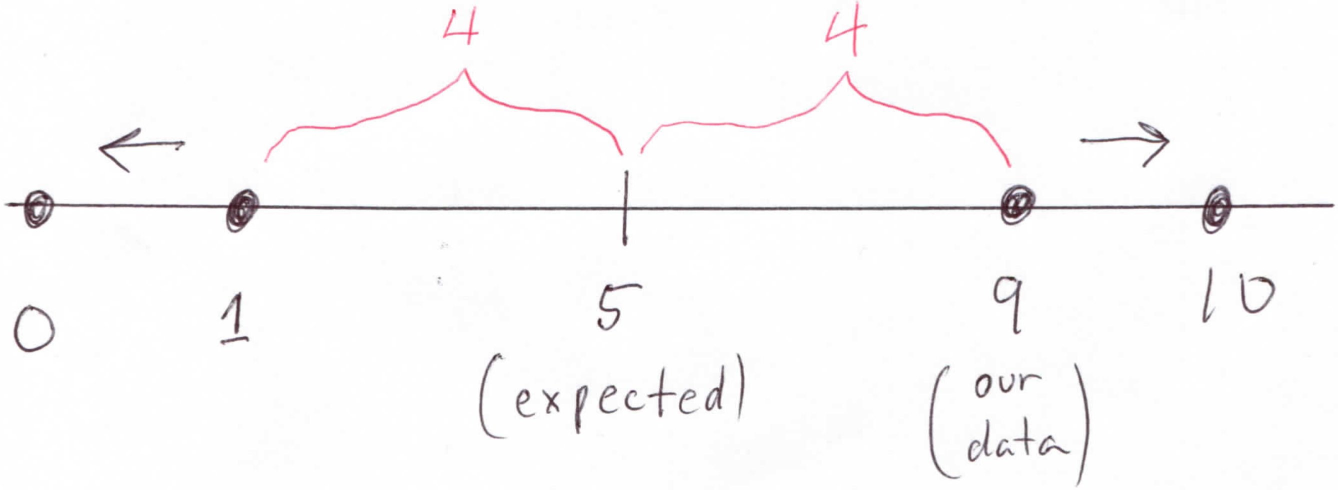
$$P(9 \text{ or more heads in ten flips of a fair coin})$$
$$= P(9 \text{ heads in ten flips of a fair coin}) + P(10 \text{ heads in ten flips of a fair coin})$$
$$= \binom{10}{9}(\frac{1}{2})^{10} + \binom{10}{10}(\frac{1}{2})^{10} = 11(\frac{1}{2})^{10} \sim 0.0107 = 1.07\%.$$

(c.) Same as (b.), except my claim now is merely that the coin of interest is not fair; that is, the $p$ of the Solution of (a.) is not 0.5.

**Solution.** As with Examples 1.7(a.), our hypothesis test is

$$H_0 : p = 0.5 \quad H_a : p \neq 0.5.$$

The outcomes relevant to our P-value now include, besides 9 or more heads, 1 or fewer heads: the numbers $0, 1, 9$, and 10 are all at least as far away from 5, the expected number of heads, as 9, the number of heads we got. Thus getting $0, 1$, or 10 heads favors $H_a$ at least as much as getting 9 heads. See the drawing directly below.




Thus our P-value now is (see APPENDIX)

$$P(0, 1, 9, \text{ or } 10 \text{ heads in ten flips of a fair coin})$$

$$= \binom{10}{0}(\frac{1}{2})^{10} + \binom{10}{1}(\frac{1}{2})^{10} + \binom{10}{9}(\frac{1}{2})^{10} + \binom{10}{10}(\frac{1}{2})^{10} = 22(\frac{1}{2})^{10} \sim 0.0215 = 2.15\%.$$

The difference between (b.) and (c.) is due to the different alternative hypotheses $H_a$. In (b.) $H_a$ only included $p > 0.5$, thus our P-value only included probabilities of *more* than 5 heads.

**2.3. Very informal definition of P-value.** P-value is measuring the *weirdness* of data, under the assumption of $H_0$. The more weird the data, when we assume $H_0$, the more we favor $H_a$, as in Definition 2.1.

In Examples 2.2, we saw that ten heads in ten flips (Examples 2.2(a.)) had a smaller P-value than nine heads in ten flips (Examples 2.2(b.)) This corresponds to ten heads in ten flips of a fair coin being weirder, that is, less likely, than nine heads in ten flips.

**Example 2.4.** I am accused of murder. As with 1.4 and 1.5, here is our hypothesis test, requiring that the accuser make a strong case, by making my innocence the default.

$H_0$: I didn't commit murder.

$H_a$: I committed murder.

Let's say there is blood on my hands soon after the fatality. Then the P-value is

$$(\text{P-value})_1 \equiv P(\text{ an innocent person having bloody hands}),$$

the probability that an innocent person has bloody hands.

Although bloody hands might make us suspicious, it is certainly not impossible for an innocent person to have bloody hands. Maybe I cut myself shaving. Thus the $(\text{P-value})_1$ is not zero.

Now suppose more data is collected: in addition to my bloody hands, I was seen attacking the victim. Now the P-value is

$$(\text{P-value})_2 \equiv P(\text{an innocent person having bloody hands being seen attacking the victim}),$$

the probability that an innocent person has bloody hands and is seen attacking the victim.

The extra data inserted into $(\text{P-value})_2$ makes an observer more suspicious of me, that is, more inclined to reject $H_0$ and convict me of murder. This increased desire to reject $H_0$ is quantified by

$$(\text{P-value})_2 < (\text{P-value})_1;$$

in general, P-value shrinkage means being more disposed to reject $H_0$.

The P-value for my null hypothesis of not committing murder shrinks as extra incriminating data is collected, corresponding, in the language of 2.3, to my appearance and activities being more weird, that is, unlikely, under the assumption of my innocence. If our data is *too* weird, that is, has a sufficiently small P-value under the null hypothesis of my innocence, we might finally give up our presumption of innocence; that is, reject $H_0$.

The nervous reader should note that, in Example 2.4, $(\text{P-value})_2$, although smaller than $(\text{P-value})_1$, is not zero. Perhaps my attack on the victim was strict Marquis of Queensberry rules, carefully limited to be nonfatal. A P-value of zero would make $H_0$ impossible; our decision making is made difficult by P-values not being zero.

**2.5. A popular misconception.** $P$-value is *not* the probability that $H_0$ is true. For example, if we flip a coin ten times and get all heads, the $P$-value for this data, relevant to the hypothesis test

$$H_0: \text{ coin is fair} \quad H_a: \text{ coin is unfair}$$

is $(\frac{1}{2})^9$, the probability of getting all heads or all tails when flipping a fair coin ten times. $(\frac{1}{2})^9$ is not the probability that the coin is fair; the coin is either fair or not fair, it is not a random event whose probability can be discussed.

See [2, Interpretation 4.4] for a similar misconception regarding confidence intervals.

## 3. TEST PROCEDURES

We saw in the last chapter that, in general, as P-values get smaller, we are more inclined to reject $H_0$; under the assumption of $H_0$, the data is increasingly weird.

We must decide how much weirdness is required to reject $H_0$; that is, how small must a P-value be for us to reject $H_0$ in favor of $H_a$.

**Definition 3.1.** Suppose $\alpha$ is a positive number less than one. A **(P-value, $\alpha$) test procedure** is the following rule for testing $H_0$ against $H_a$, as in 1.8.

$$\text{P-value} \leq \alpha \;\text{ implies you reject } H_0;$$

and

$$\text{P-value} > \alpha \;\text{ implies you do not reject } H_0.$$

The number $\alpha$ will be seen (Definitions 3.4 and Theorem 3.6) to be fundamental in other ways besides its definition as the minimum weirdness required for rejection of $H_0$.

**Definitions 3.2.** Much more generally than 3.1, a **test procedure** for testing $H_0$ against $H_a$ spells out, in advance of collecting data, which data will cause you to reject $H_0$. The set of all such data is a **rejection region**, succinctly describing the test procedure.

For a (P-value, $\alpha$) test procedure (Definition 3.1) the rejection region is

$$\{\text{data whose P-value is } \leq \alpha\}.$$

**Example 3.3.** In the hypothesis test in Examples 2.2(c.), we could make our rejection region be getting all heads or all tails, when flipping the coin ten times. The test procedure is then

$$\text{reject } H_0 \text{ if you get all heads or all tails}$$

and

$$\text{do not reject } H_0 \text{ if you get neither all heads nor all tails}$$

**Definitions 3.4.** The **significance level** of a test procedure is

$$\alpha \equiv P\,(\text{reject } H_0 \text{ when } H_0 \text{ is true}) = P\,(\text{rejection region})\,, \text{under the assumption of } H_0.$$

The set that we just took a probability of is called a **Type I error**. An example of a Type I error is convicting me of murder when I'm innocent (at least of murder). With our reluctance to reject $H_0$, Type I error is much more serious than **Type II error**, meaning failing to reject $H_0$ when $H_0$ is false. An example of Type II error is a murderer being acquitted in court.

Our caution about rejecting $H_0$, as in 1.5 and 1.8, implies that we want $\alpha$ to be small.

Notice that, for the test procedure in Definition 3.1, the P-value is the smallest significance level at which we would reject $H_0$. This explains the terminology "observed significance level" for P-value, in Definition 2.1.

**Example 3.5.** The significance level of the test procedure in Example 3.3 is

$$\alpha \equiv P\,(\text{getting ten heads or ten tails when flipping a fair coin}) = 2(\frac{1}{2})^{10} = \frac{1}{2^9} \sim 0.002.$$

Still addressing the test procedure in Example 3.3: if we got nine heads and one tail, we would not reject $H_0$; we would say there is insufficient evidence, at significance level $\frac{1}{2^9}$, to conclude the coin is unfair. We would *not* conclude the coin is fair (see 1.8).

Note that the test procedure in Example 3.3 is a (P-value, $\frac{1}{2^9}$) test procedure (see Definition 3.1), since the only data that has a P-value less than or equal to $\frac{1}{2^9}$ is having all heads or all tails (see HW19).

For the following result relating P-values to significance levels we choose to not state what the "large class" is, nor prove our result. Theorem 3.6 finishes the discussion we started at the beginning of this chapter. We will see an example of Theorem 3.6 (Theorem 4.7) in the next chapter. All the test procedures of interest that this author knows about satisfy Theorem 3.6; that is, are as in Definition 3.1.

**Theorem 3.6.** See Definition 3.1.

(a) For a large class of P-value assignments and positive numbers $\alpha$ less than one, the (P-value, $\alpha$) test procedure has significance level $\alpha$.

(b) A large class of test procedures are (P-value, $\alpha$) test procedures, with $\alpha$ equal to the significance level of the procedure.

**Examples 3.7.** (a.) This is an example of a hypothesis test and P-value assignments for which Theorem 3.6(a) is not true.

Let our hypothesis test be as in Examples 1.7(a.) and 2.2(c.), with the same data collection: flip the coin of interest ten times and count the number of heads.

Take the $\alpha$ of Theorem 3.6 to be 0.01.

We will leave it to the reader (see HW19) to show that

$$\text{P-value} \leq 0.01$$

if and only if the ten flips are either all heads or all tails.

It follows that the rejection region is zero heads or ten heads, as in Example 3.3, so that the significance level of our test procedure is, as calculated in Example 3.5, $\frac{1}{2^9} \sim 0.002$, *not* the $\alpha$ of 0.01 we started with.

(b.) Now we'll give an example where Theorem 3.6(b) fails.

Take the same hypothesis test and data collection as in (a.), and let the rejection region for our test procedure be getting precisely one head in ten flips of a fair coin.

Our significance level is

$$\alpha = P(\text{one head in ten flips of a fair coin}) = \frac{10}{2^{10}} \text{ (see APPENDIX)},$$

thus

$$\text{P-value} \leq \alpha$$

if and only if we got all heads or all tails (see HW19), a different rejection region than our test procedure, so that our test procedure is *not* a (P-value, $\alpha$) test procedure.

**Discussion 3.8.** Making the significance level $\alpha$ smaller means we require more weirdness (smaller P-values) to reject $H_0$. See the next chapter for examples.

There are many possible factors that might make one choose smaller significance levels, to make it harder (that is, require more compelling data) to reject $H_0$. In Example 2.4, where $H_0$ is me being innocent of murder, the presence of the death penalty for murder might motivate a smaller significance level, to make it harder to convict me of murder.

**Remarks 3.9.** The advantage of focusing on rejection regions for a test procedure is that we don't need to calculate P-values, with their possible ambiguities. The disadvantage of rejection regions is that we can construct them only one $\alpha$ at a time, and receive only "yes, reject $H_0$," or "no, don't reject $H_0$" as a conclusion; this is much less information than a P-value, with its continuum of possible information.

## 4. HYPOTHESIS TESTING FOR A (special case of a) POPULATION MEAN

Throughout this chapter $\mu$ is the (unknown) mean of a normal random variable $X$ with known standard deviation $\sigma$, $\mu_0$ is a number and $n$ is the sample size of a random sample.

See [2, Chapter 2] for needed information about normal populations, especially 2.3, 2.5, and 2.7; see also [3] for basic statistical terminology, especially 6, 10, and 17.

Our null hypothesis throughout this chapter is

$$H_0 : \mu = \mu_0,$$

for $\mu_0$ to be specified. What we need for hypothesis testing is summarized in Tables 4.2 and 4.5, and Theorem 4.7.

**Discussion 4.1.** Given a random sample of measurements $x_1, x_2, \ldots, x_n$ from $X$, to calculate P-value as in 2.1 or 2.3, the measurement that first springs to mind is the sample mean (see [3, Definitions 6]).

$$\overline{x} \equiv \frac{1}{n}(x_1 + x_2 + \cdots + x_n);$$

intuitively, the further $\overline{x}$ is from $\mu_0$, the more dubious we are about $H_0$; that is, the more we favor the alternative hypothesis $H_a$ over $H_0$, as in 2.1. In the language of 2.3, the weirdness of our data could be plausibly measured by the distance from $\overline{x}$ to $\mu_0$.

But we can improve our strategy by using the fact (see [2, Theorem 2.7]) that

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}},$$

unleashing the power of $Z$ tables, such as those attached to the end of this magnification.

This means that we will measure weirdness of data by looking at the distance from

$$z \equiv \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \quad \text{to} \quad \frac{\mu_0 - \mu_0}{\frac{\sigma}{\sqrt{n}}} = 0;$$

that is, the absolute value of $z$; the larger $|z|$ is, the more we are inclined to reject $H_0$ in favor of the alternative hypothesis $H_a$.

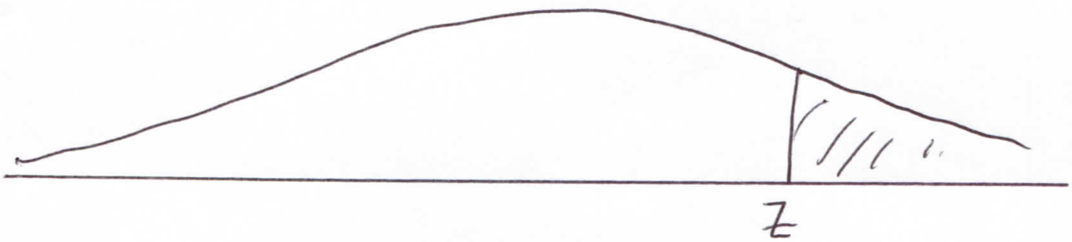
Both the random variable $Z = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ and the processed data $z = \frac{\overline{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ are denoted *test statistics* we will use for this hypothesis test.

Our precise P-value definitions for the null hypothesis $H_0$ of this chapter are different for different alternative hypotheses $H_a$, as summarized in the table on the next page.

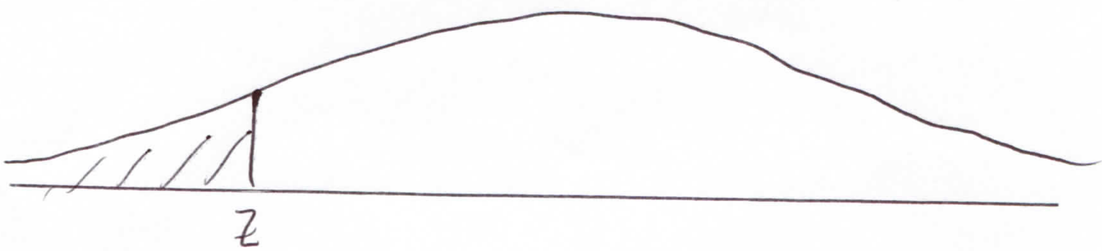
### 4.2. HYPOTHESIS TEST P-values for $H_0 : \mu = \mu_0$, after calculating test statistic $z = \frac{\overline{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$

| $H_a$ | $P$-value |
|---|---|
| (1) $\mu > \mu_0$ | $P(Z > z)$ |
| (2) $\mu < \mu_0$ | $P(Z < z)$ |
| (3) $\mu \neq \mu_0$ | $[P(Z > \|z\|) + P(Z < -\|z\|)]$ |
| | $= 2P(Z > \|z\|)$ |

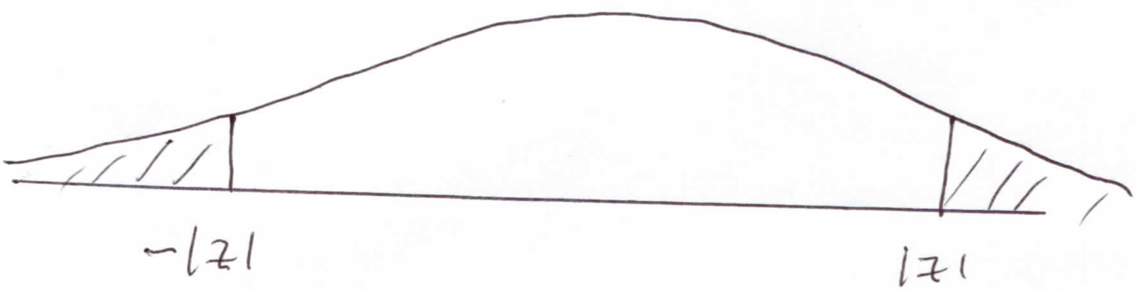
(1)




(2)




(3)



**Example 4.3.** Suppose $X$ is normal, with standard deviation $\sigma = 5$. For the hypothesis test

$$H_0 : \mu = 2 \quad H_a : \mu < 2,$$

find the P-value of the following measurements of $X$:

$$x_1 = 0, x_2 = 10, x_3 = 2, x_4 = -8.$$

**Solution.** We need

$$\overline{x} = \frac{1}{4}(0 + 10 + 2 - 8) = 1.$$

We also have $\sigma = 5, \mu_0 = 2$, and sample size $n = 4$, thus our test statistic is

$$z = \frac{(\overline{x} - 2)}{\frac{5}{\sqrt{4}}} = \frac{(1 - 2)}{\frac{5}{\sqrt{4}}} = -0.4.$$

Since $H_a$ is of the form (2) in 4.2, measure weirdness, as in 2.3, with values of $Z$ less than our calculated test statistic $z$; that is, from (2) of 4.2, our P-value is

$$P(Z < -0.4) = P(Z > 0.4) = 0.3446.$$

**Discussion 4.4.** Let's translate the P-values of 4.2 into a test procedure as in 3.1 and 3.2.

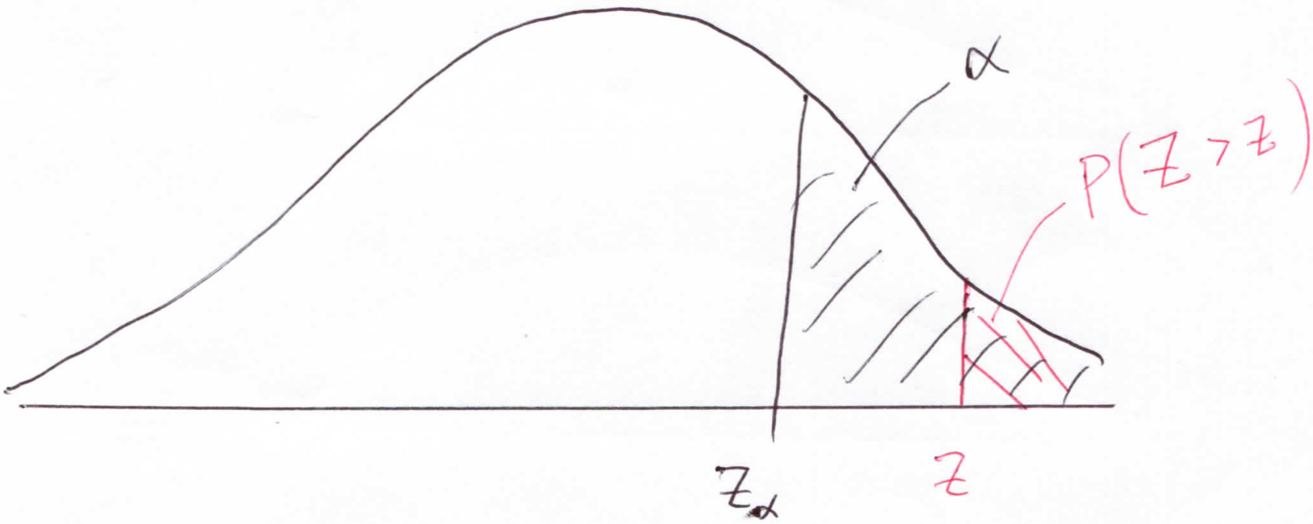
Given positive $\alpha$ less than one, for $H_a : \mu > \mu_0$, 3.1 tells us to reject $H_0$ if and only if

$$P(Z > z) = P\text{-value} \leq \alpha = P(Z > z_\alpha),$$

which is equivalent to $z \geq z_\alpha$, the *critical value* (see [2, Definition 2.5]).

In the language of 3.2, $z \geq z_\alpha$ is the *rejection region* for the test procedure of 3.1.

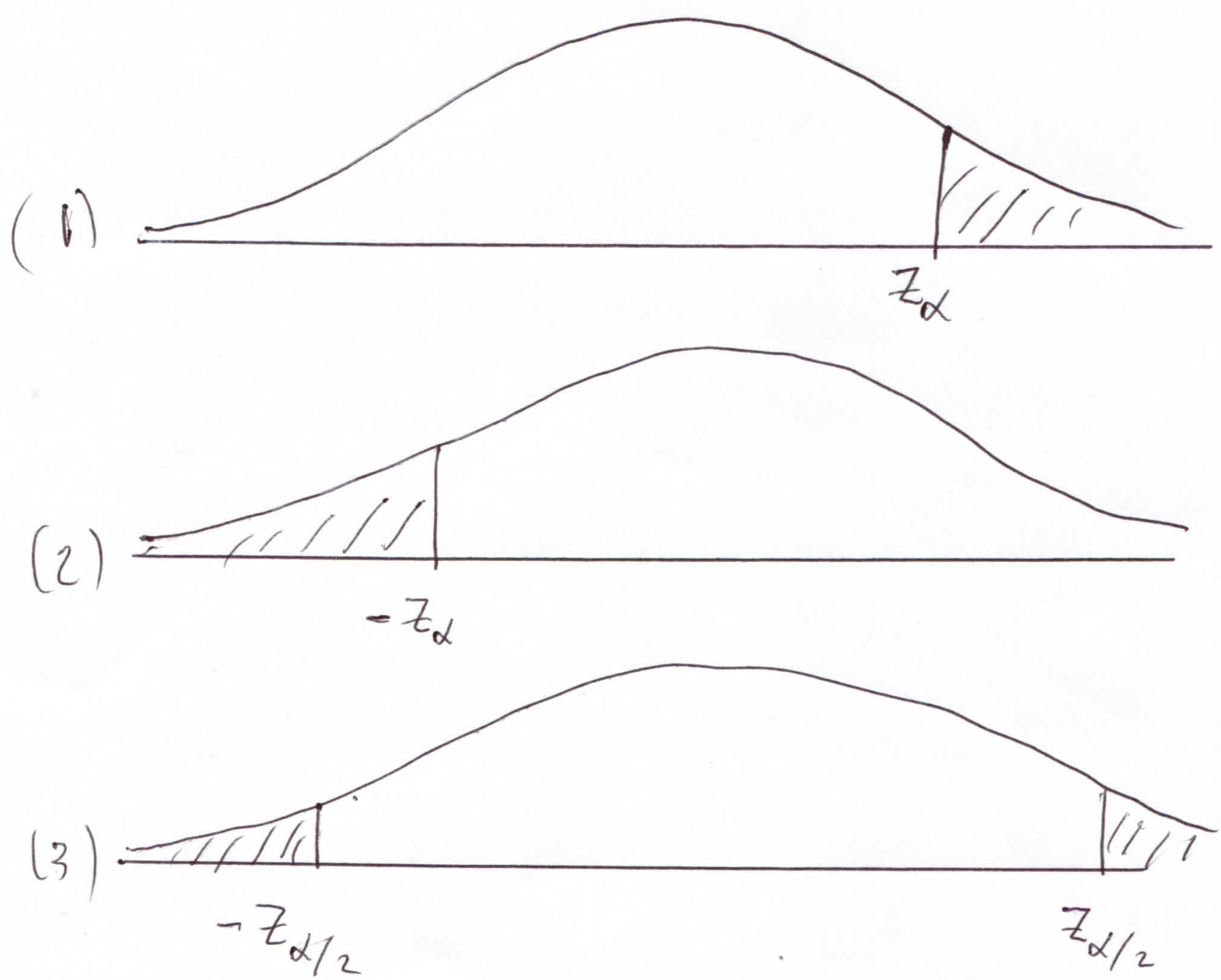
See the drawing directly below, where $\alpha = P(Z > z_\alpha)$ is the black-shaded area and the P-value $P(Z > z)$ is the red-shaded area.




Similar reasoning for the other alternative hypotheses gives us the test procedure summarized in the table on the next page.

**4.5. HYPOTHESIS TEST Rejection Regions for $H_0 : \mu = \mu_0$, after calculating test statistic $z = \frac{\overline{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$**

| $H_a$ | Rejection Region |
|---|---|
| (1) $\mu > \mu_0$ | $z \geq z_\alpha$ |
| (2) $\mu < \mu_0$ | $z \leq -z_\alpha$ |
| (3) $\mu \neq \mu_0$ | $\lvert z \rvert \geq z_{\frac{\alpha}{2}}$ |



**Example 4.6.** Suppose $X$ is normal, with standard deviation $\sigma = 5$. Test

$$H_0 : \mu = 2 \quad H_a : \mu < 2,$$

at significance level 0.05, if we have made the following measurements of $X$:

$$x_1 = 0, x_2 = 10, x_3 = 2, x_4 = -8.$$

**Solution.** In Example 4.3 we got the test statistic $z = -0.4$.
From the $Z$ tables we get $z_{0.05} = 1.65$. Focusing on (2) of 4.5, we compare $z$ to $-z_{0.05}$:

$$z = -0.4 > -1.65 = -z_{0.05}$$

implies that $z$ is *not* in the rejection region, thus we don't reject $H_0$. At significance level 0.05, there is insufficient evidence to conclude that $\mu < 2$.

The following should be compared to 3.6 and 3.1.

**Theorem 4.7.** Suppose $\alpha$ is a positive number less than one. Then the test procedure given by 4.5 has significance level $\alpha$, and is equivalent to

$$\text{P-value} \leq \alpha \text{ implies you reject } H_0;$$

and

$$\text{P-value} > \alpha \text{ implies you do not reject } H_0,$$

for the P-values defined by 4.2.

**Proof:** The equivalence of the rejection regions in 4.5 and the P-value comparisons we proved in Discussion 4.4, at least for $H_a : \mu > \mu_0$; almost identical arguments show the same equivalence for the other choices of $H_a$. The significance level of $\alpha$ is merely the definitions of significance level (Definitions 3.4) and of critical value $z_\alpha$([2, Definition 2.5]). □

**Example 4.8.** Here is Example 4.6 done with P-values.

In Example 4.3, we calculated the P-value of our data to be 0.3446.
Since

$$\text{P-value} = 0.3446 > 0.05 = \alpha,$$

we don't reject $H_0$, at significance level $\alpha$, by Theorem 4.7.

**More Examples 4.9.** (a.) Same as Example 4.6, except we have 100 measurements with an average of 1.

**Solution.** Since $n$ is now 100, we have a new test statistic

$$z = \frac{(\overline{x} - \mu_0)}{\frac{\sigma}{\sqrt{n}}} = \frac{(1-2)}{\frac{5}{\sqrt{100}}} = -2.$$

Let's perform this hypothesis test in two ways, first with a P-value, then with a rejection region.
With P-value (see 4.2(2)):

$$\text{P-value} = P(Z < -2) = P(Z > 2) = 0.0228 \leq 0.05 \equiv \alpha,$$

thus we reject $H_0$; there is sufficient evidence to conclude, at significance level 0.05, that $\mu < 2$.

With rejection region (see 4.5(2)):

$$z_\alpha = z_{0.05} = 1.65,$$

thus our rejection region is

$$z \leq -1.65.$$

Since our test statistic is

$$z = -2 \leq -1.65,$$

our test statistic is in the rejection region, thus we reject $H_0$, as stated in the conclusion arrived at with a P-value.

Notice that the same discrepancy, between the sample mean $\overline{x} = 1$ and the hypothesized population mean $\mu_0 = 2$ led to different conclusions, in Example 4.6 compared to Examples 4.9(a.), only because Examples 4.9(a.) had a larger sample size.

The intuition here is that a larger sample means more information, hence the ability to draw more conclusions; in this case, the decision to reject $H_0$.

(b.) Same as (a.), except $H_a : \mu \neq 2$.

**Solution.** Working with P-values as in 4.2(3), we have, since our test statistic is unchanged from (a.),

$$\text{P-value} = P(Z < -2) + P(Z > 2) = 2P(Z > 2) = 2(0.0228) = 0.0456 \leq 0.05 \equiv \alpha,$$

thus we reject $H_0$; there is sufficient evidence to conclude, at significance level 0.05, that $\mu \neq 2$.

Notice that, since the alternative hypothesis $H_a$ allows values of $\mu$ larger or smaller than $\mu_0 = 2$, our P-value, measuring weirdness, allows both large and small values of $\overline{x}$, hence $z$, to contribute to weirdness.

If we had done this problem with rejection regions, we would need

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96,$$

producing (see 4.5(3)) a rejection region of

$$|z| \geq 1.96.$$

Our test statistic of $z = -2 \leq -1.96$ is only *barely* in the rejection region. Merely saying "yes, reject" leaves out a lot of information.

(c.) Same as (a.), except at significance level 0.01.

**Solution.** We calculated in (a.) the P-value of 0.0228. Since this is greater than the significance level $\alpha \equiv 0.01$, we do not reject $H_0$; there is insufficient evidence to conclude, at significance level 0.01, that $\mu < 2$.

The same failure to reject is true for any significance level $\alpha < 0.0228$, our P-value (see Theorem 4.7).

(d.) Test, at significance level 0.05, the claim that fish makes you smarter, on average, if the IQs of nine fish-eating people are

$$97, 121, 89, 100, 100, 134, 70, 128, 97.$$

Assume that IQ has a mean of 100, and the IQ of fish eaters is normally distributed, with a standard deviation of 15.

**Solution.** Let $\mu$ be the average IQ of fish eaters. Here is our hypothesis test.

$$H_0 : \mu = 100 \quad H_a : \mu > 100.$$

For either the P-value (4.2(1)) or the rejection region (4.5(1)) approach, we need the sample mean

$$\overline{x} = \frac{1}{9}(97 + 121 + 89 + 100 + 100 + 134 + 70 + 128 + 97) = 104.$$

We're given $n = 9$ and $\sigma = 15$, so our test statistic is

$$z = \frac{(104 - 100)}{\frac{15}{\sqrt{9}}} = 0.8.$$

Our P-value is

$$P(Z > 0.8) = 0.2119 > 0.05 \equiv \alpha,$$

thus (see Theorem 4.7) we do not reject $H_0$; the evidence is not sufficient, at significance level 0.05, to conclude that fish makes you smarter.

Using rejection regions, we would first get the critical value $z_\alpha = z_{0.05} = 1.65$, so that

$$z = 0.8 < 1.65 = z_\alpha;$$

that is, our test statistic $z$ is not in the rejection region for $H_0$ (see 4.5(1)), thus we do not reject $H_0$.

(e.) Same as (d.), except our data is 81 fish-eating people, with an average IQ of 104.

**Solution.** Our test statistic is now

$$z = \frac{(104 - 100)}{\frac{15}{\sqrt{81}}} = 2.4.$$

Using rejection regions, we now have

$$z = 2.4 \geq 1.65 = z_\alpha,$$

so that $z$ is in the rejection region for $H_0$: the evidence is now sufficient to conclude, at significance level 0.05, that fish makes you smarter (on average).

Using P-values, we would calculate

$$\text{P-value} = P(Z > 2.4) = 0.0082 \leq 0.05 \equiv \alpha,$$

thus we reject $H_0$.

(f.) Same as (e.), except significance level 0.005.

**Solution.** We've already calculated the P-value:

$$\text{P-value} = P(Z > 2.4) = 0.0082 > 0.005 \equiv \alpha,$$

thus we do not reject $H_0$; at significance level 0.005, the evidence is insufficient to conclude that fish makes you smarter.

(g.) Suppose 16 people taking a pill advertised to be a diet pill have the following weights, in pounds:

$$130, 100, 160, 150, 110, 100, 160, 110, 150, 130, 120, 120, 150, 100, 140, 150.$$

Does this data provide strong evidence (in this case, significance level 0.01) that the advertised pill makes people lose weight (on average)? Assume the mean weight of randomly chosen people is 160 pounds and the weights of people taking this pill are normally distributed, with a standard deviation of 20 pounds.

**Solution.** Let $\mu$ be the average weight, in pounds, of all people taking this diet pill. Here is our hypothesis test.

$$H_0 : \mu = 160 \quad H_a : \mu < 160.$$

We'll need the sample mean $\overline{x}$, calculated, as in (d.), to be 130. Since $n = 16$ and $\sigma = 20$, our test statistic is

$$z = \frac{(130 - 160)}{\frac{20}{\sqrt{16}}} = -6.$$

Getting a rejection region is easier than working with P-values, since $|z|$ is larger than 3:

$$z = -6 \leq -2.33 = z_{0.01},$$

thus $z$ is in the rejection region for $H_0$; at significance level 0.01, the evidence is sufficiently strong to assert that the advertised pill makes people lose weight (on average).

(h.) Same as (g.), except advertisements have been escalated to losing at least twenty pounds; same data as in (g.).

**Solution.** Since the advertisers now want the average weight of pill takers to be less than 140, the hypothesis test is modified to

$$H_0 : \mu = 140 \quad H_a : \mu < 140.$$

Our test statistic is now

$$z = \frac{(130 - 140)}{\frac{20}{\sqrt{16}}} = -2,$$

so that

$$z = -2 > -2.33 = -z_{0.01},$$

not in the rejection region for $H_0$, so the data is not providing strong enough evidence to conclude that the pill makes people lose at least twenty pounds (on average).

(i.) Suppose the average length of sleep at night, in hours, of 64 coffee drinkers, is 7.5 hours.

Does the data suggest, at significance level 0.005, that coffee changes how long you sleep at night? Assume that length of sleep at night of a randomly chosen person has a mean of 8 hours, and the length of sleep of coffee drinkers is normally distributed, with a standard deviation of 1.5 hours.

**Solution.** Let $\mu$ be the average number of hours slept, among all coffee drinkers. Our hypothesis test is

$$H_0 : \mu = 8 \quad H_a : \mu \neq 8.$$

We have $\overline{x} = 7.5, n = 64$, and $\sigma = 1.5$, so our test statistic is

$$z = \frac{(7.5 - 8)}{\frac{1.5}{\sqrt{64}}} = -\frac{8}{3} \sim -2.67,$$

giving us a P-value of

$$P(|Z| > 2.67) = 2P(Z > 2.67) = 2 \times 0.0038 = 0.0076 > 0.005 \equiv \alpha,$$

thus we do not reject $H_0$; at significance level 0.005, the data does not suggest that coffee changes how long you sleep at night.

(j.) Same as (i.), except claim to be tested is "coffee makes you sleep less at night."

**Solution.** Now we are testing

$$H_0 : \mu = 8 \quad H_a : \mu < 8.$$

We have the same test statistic $z \sim -2.67$, as in (i), but the P-value (see 4.2) changes, because $H_a$ has changed:

$$\text{P-value} = P(Z < -2.67) = P(Z > 2.67) = 0.0038 \leq 0.005 \equiv \alpha,$$

so that we now reject $H_0$: the data suggests, at significance level 0.005, that (on average) coffee makes you sleep less at night.

The intuition, in (i.) versus (j.), is that (j.) has more implicit information, because of its more restrictive $H_a$; in general, increased information makes it more possible to reject null hypotheses. See problems 14 versus 15, in the homework for this magnification.

(k.) Suppose all we know about the sample mean $\overline{x}$ is that it's greater than $(\mu_0 + 2)$. How large must the sample size be, so that we reject $H_0$ at significance level 0.002, in the hypothesis test

$$H_0 : \mu = \mu_0 \quad H_a : \mu > \mu_0?$$

Assume $\mu$ is the (unknown) mean of a normal population with standard deviation 20.

**Solution.** Denoting by $n$ the sample size, let's say as much as possible about the test statistic

$$z = \frac{(\overline{x} - \mu_0)}{\frac{20}{\sqrt{n}}} > \frac{2}{\frac{20}{\sqrt{n}}} = (0.1)\sqrt{n}.$$

We need the critical value

$$z_\alpha = z_{0.002} = 2.88.$$

Rejection of $H_0$ occurs when $z \geq z_\alpha$. Specifically we want $(0.1)\sqrt{n} \geq 2.88$; solving for $n$ gives us $n \geq 829.44$. We want a sample size greater than 829.

# 5. HYPOTHESIS TESTING and CONFIDENCE INTERVALS or BOUNDS

As in Chapter 4, throughout this chapter $\mu$ is the mean of a normal random variable $X$ with known standard deviation $\sigma$, $\mu_0$ is a number and $n$ is the sample size of a random sample.

This chapter shows the equivalence of the two most popular forms of statistical inference, hypothesis testing and confidence intervals or bounds ([2, especially Chapter 3]), at least for inference on the population mean of a normal random variable with known standard deviation. A future magnification will give a much wider class of inference where this equivalence remains true.

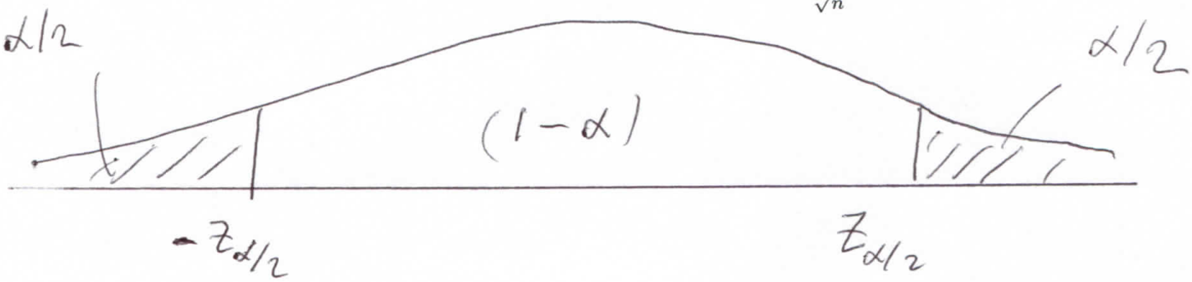
**Discussion 5.1.** We shall see (Theorem 5.2) that confidence intervals correspond to the two-sided alternative hypothesis $H_a : \mu \neq \mu_0$, upper confidence bounds to the one-sided alternative hypothesis $H_a : \mu < \mu_0$, and lower confidence bounds to the one-sided alternative hypothesis $H_a : \mu > \mu_0$. This discussion will restrict itself to the first correspondence we just listed.

Both CIs and hypothesis tests in the setting of this chapter begin with (see [2, Chapter 3]) the test statistic

$$(*) \quad Z = \frac{(\overline{X} - \mu)}{\frac{\sigma}{\sqrt{n}}},$$

so that

$$(**) \quad (1-\alpha) = P(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) = P(-z_{\frac{\alpha}{2}} < \frac{(\overline{X} - \mu)}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}).$$




Solving for $\mu$ as in [2, 3.1] gives us

$$(1-\alpha) = P(\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}),$$

so that our formula for a $100(1-\alpha)\%$ *confidence interval* for $\mu$ is

$$\left(\left(\overline{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right), \left(\overline{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)\right) \equiv \text{the set of all real numbers } c \text{ satisfying } \left(\overline{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) < c < \left(\overline{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right);$$

this is also denoted

$$\overline{x} \pm \left(z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right).$$

See [2, Definitions 3.2]

We could similarly solve for $\overline{X}$ in (**), to aim for the rejection region for a test procedure. But we find it convenient, when $\mu = \mu_0$ as in our null hypothesis in Chapters 4 and 5, to use the test statistic (*) above, so that (**) becomes

$$(1-\alpha) = P(|Z| < z_{\frac{\alpha}{2}})$$

or, looking at complements,

$$\alpha = P(|Z| \geq z_{\frac{\alpha}{2}}).$$

The shaded area in (**) becomes the rejection region for $H_0$, so that $\alpha$, its area, is the *significance level* of the test procedure (see Definitions 3.4 and Table 4.5(3)).

Note that the confidence level of $(1-\alpha)$ in the confidence interval corresponds to the significance level of $\alpha$ in the hypothesis test

$$H_0 : \mu = \mu_0 \;\; H_a : \mu \neq \mu_0.$$

Here is the precise statement of the relationship between hypothesis tests and confidence intervals or upper or lower confidence bounds. See the drawings on the next page.

**Theorem 5.2.** Suppose $\alpha$ is a positive number less than one.

(a) Reject $H_0$, at significance level $\alpha$, in

$$H_0 : \mu = \mu_0 \;\; H_a : \mu > \mu_0$$

if and only if $\mu_0$ is less than or equal to the $100(1-\alpha)\%$ lower confidence bound for $\mu$.

(b) Reject $H_0$, at significance level $\alpha$, in

$$H_0 : \mu = \mu_0 \;\; H_a : \mu < \mu_0$$

if and only if $\mu_0$ is greater than or equal to the $100(1-\alpha)\%$ upper confidence bound for $\mu$.

(c) Reject $H_0$, at significance level $\alpha$, in

$$H_0 : \mu = \mu_0 \;\; H_a : \mu \neq \mu_0$$

if and only if $\mu_0$ is not in the $100(1-\alpha)\%$ confidence interval for $\mu$.

**Proof:** See 4.5 and [2, Definitions 3.2].

(a) Don't reject $H_0$, at significance level $\alpha$, in

$$H_0 : \mu = \mu_0 \;\; H_a : \mu > \mu_0$$

$\iff$

$$\frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < z_\alpha \iff \mu_0 - \overline{x} > -z_\alpha \frac{\sigma}{\sqrt{n}} \iff \mu_0 > \left(\overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}\right),$$

so that we

$$\text{reject } H_0 \iff \mu_0 \leq \left(\overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}\right);$$

since $\left(\overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}\right)$ is the $100(1-\alpha)\%$ lower confidence bound for $\mu$, this concludes the proof of (a).

(b) is very similar to (a) and is left to the reader.

(c) Don't reject $H_0$, at significance level $\alpha$, in

$$H_0 : \mu = \mu_0 \;\; H_a : \mu \neq \mu_0$$

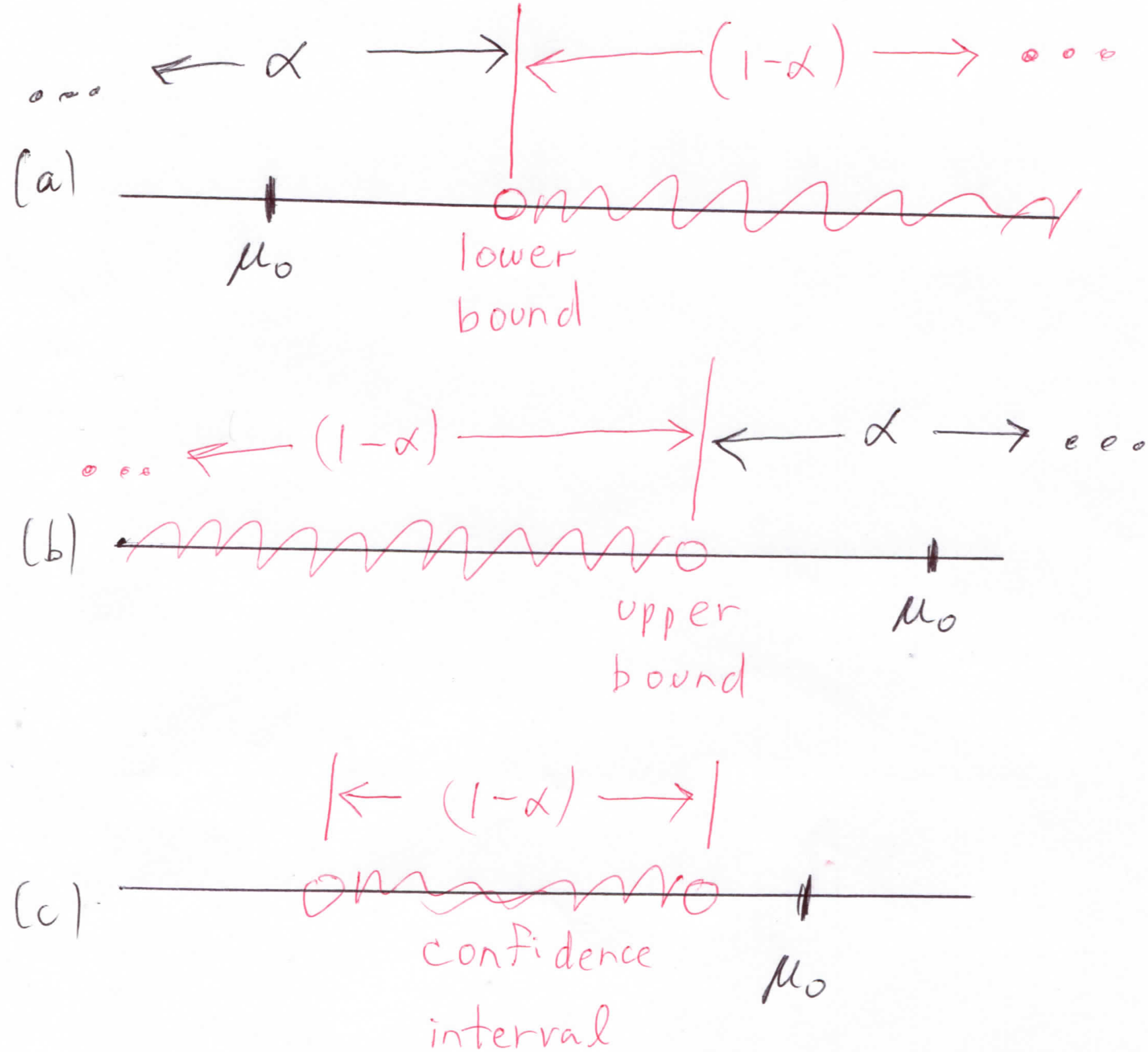
$\iff$

$$\left|\frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right| < z_{\frac{\alpha}{2}} \iff |\overline{x} - \mu_0| < z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \iff |\mu_0 - \overline{x}| < z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\iff -z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < (\mu_0 - \overline{x}) < z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \iff \left(\overline{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) < \mu_0 < \left(\overline{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$$

$\iff \mu_0$ is in the $100(1-\alpha)\%$ confidence interval for $\mu$. $\square$

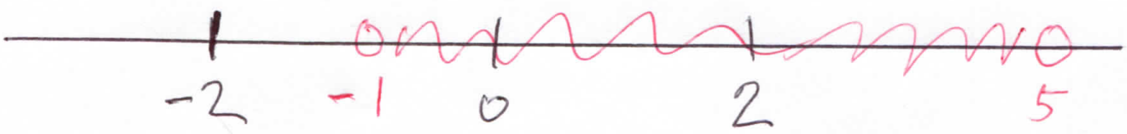
**Theorem 5.2 drawn**



**Examples 5.3.** (a.) Suppose data produces the 90% confidence interval $(-1, 5)$ for $\mu$. Test each of the following, at significance level 0.1.

$$(i)\ H_0 : \mu = 0 \quad H_a : \mu \neq 0$$
$$(ii)\ H_0 : \mu = 2 \quad H_a : \mu \neq 2$$
$$(iii)\ H_0 : \mu = -2 \quad H_a : \mu \neq -2$$

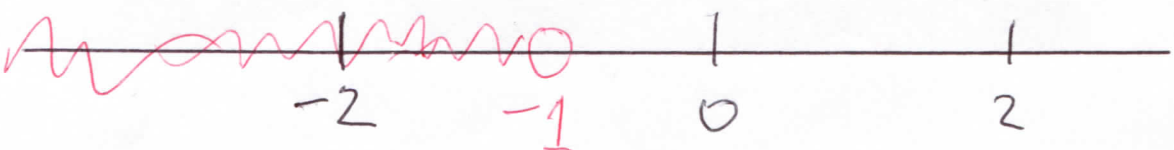
**Solutions.** See Theorem 5.2(c). Since 0 and 2 are in $(-1, 5)$, we don't reject $H_0$ in (i) and (ii); since $-2$ is not in $(-1, 5)$, we reject $H_0$ in (iii).




(b.) Suppose data produces $-1$ as a 90% upper confidence bound for $\mu$. Test each of the following, at significance level 0.1.

$$(i)\ H_0 : \mu = 0 \quad H_a : \mu < 0$$
$$(ii)\ H_0 : \mu = 2 \quad H_a : \mu < 2$$
$$(iii)\ H_0 : \mu = -2 \quad H_a : \mu < -2$$

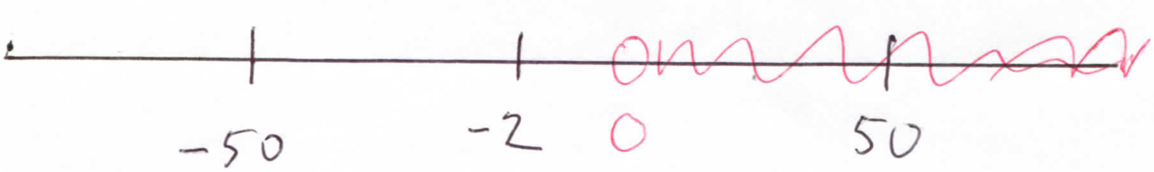
**Solutions.** See Theorem 5.2(b). Since 0 and 2 are greater than or equal to $-1$, we reject $H_0$ in (i) and (ii); since $-2$ is less than $-1$, we don't reject $H_0$ in (iii).




(c.) Suppose data produces 0 as a 99% lower confidence bound for $\mu$. Test each of the following, at significance level 0.01.

$$(i)\ H_0 : \mu = 50 \quad H_a : \mu > 50$$
$$(ii)\ H_0 : \mu = -50 \quad H_a : \mu > -50$$
$$(iii)\ H_0 : \mu = -2 \quad H_a : \mu > -2$$

**Solutions.** See Theorem 5.2(a). Since 50 is greater than 0, we don't reject $H_0$ in (i); since $-50$ and $-2$ are less than or equal to 0, we reject $H_0$ in (ii) and (iii).



## APPENDIX: FLIPPING A FAIR COIN

See [1] for more on the subject of this appendix.

A coin is **fair** if, on each flip, the probability of getting heads equals the probability of getting tails.

To address probabilities of getting a specified number of heads when flipping a fair coin a specified number of times, we will find the following terminology very useful.

**Definitions APP.1. zero factorial**, denoted $0!$, is defined to be 1.

**one factorial**, denoted $1!$, is defined to be 1.

**two factorial**, denoted $2!$, is defined to be $2 \times 1 = 2$.

**three factorial**, denoted $3!$, is defined to be $3 \times 2 \times 1 = 6$.

In general, for $n$ a natural number,

**n factorial**, denoted $n!$, is defined to be $n \times (n-1) \times (n-2) \cdots \times 3 \times 2 \times 1$.

Factorials grow very quickly: $4! = 24, 5! = 120, 6! = 720, 7! = 5,040$, etc.

The following counting tells you, for $n, k = 0, 1, 2, \ldots, k \leq n$, how many subsets of size $k$ may be chosen from a set of size $n$; for example, the number of poker hands (five cards) you can get from a deck of fifty-two cards.

**Definition APP.2.** For $n, k$ natural numbers, with $n \geq k$, **n choose k**, denoted $\binom{n}{k}$, is

$$\frac{n!}{k!(n-k)!} = \frac{n \times (n-1) \times (n-2) \cdots (n-k+1)}{k!}.$$

**Examples APP.3.** $\binom{12}{2} = \frac{12!}{2!10!} = \frac{12 \times 11}{2!} = 66$.

$\binom{9}{4} = \frac{9!}{4!5!} = \frac{9 \times 8 \times 7 \times 6}{4!} = 126$.

The number of poker hands is $\binom{52}{5} = \frac{52!}{5!47!} = \frac{52 \times 51 \times 50 \times 49 \times 48}{5!} = 2,598,960$.

**Theorem APP.4.** For $n, k$ natural numbers, with $n \geq k$,

$$P(k \text{ heads in } n \text{ flips of a fair coin}) = \left(\frac{1}{2}\right)^n \binom{n}{k}.$$

**Examples APP.5.** The probability of getting four heads in ten flips of a fair coin is

$$\left(\frac{1}{2}\right)^{10} \binom{10}{4} = \frac{210}{2^{10}} = \frac{105}{512}.$$

The probability of getting three heads in six flips of a fair coin is

$$\left(\frac{1}{2}\right)^{6} \binom{6}{3} = \frac{20}{2^{6}} = \frac{5}{16}.$$

# HOMEWORK

Assume, in all problems, that P-values and test procedures satisfy Theorem 3.6.

**1.** Set up each of the hypothesis tests in 1.6 in the form, for some parameter $\theta$, number $\theta_0$,

$$H_0 : \theta = \theta_0 \quad H_a : \theta > \theta_0$$

or

$$H_0 : \theta = \theta_0 \quad H_a : \theta < \theta_0$$

or

$$H_0 : \theta = \theta_0 \quad H_a : \theta \neq \theta_0.$$

See 1.7.

**2.** Set up each of the following claims:

(a) vaccinations will make people lose weight

(b) vaccinations will change people's weight

and

(c) Brutopia is stronger than Fredonia

in the form, for some population mean $\mu$, number $\mu_0$,

$$(i) \quad H_0 : \mu = \mu_0 \quad H_a : \mu > \mu_0$$

or

$$(ii) \quad H_0 : \mu = \mu_0 \quad H_a : \mu < \mu_0$$

or

$$(iii) \quad H_0 : \mu = \mu_0 \quad H_a : \mu \neq \mu_0.$$

See 1.7.

**3.** Which of the following (choose one of (a) or (b)) has a smaller P-value, for the null hypothesis of my not being a counterfeiter?

(a) I have 100 identical counterfeit five-dollar bills.

(b) I have 100 identical counterfeit five-dollar bills and a printing press that specializes in printing objects the size of a five-dollar bill.

**4.** Suppose I reject a null hypothesis at significance level 0.01.

(a) Would I automatically also reject that null hypothesis at significance level 0.05?

(b) Would I automatically also reject that null hypothesis at significance level 0.001?

**5.** Suppose the P-value for a hypothesis test with null hypothesis $H_0$ is 0.035.

(a) Would I reject $H_0$ at significance level 0.01?

(b) Would I reject $H_0$ at significance level 0.05?

(c) What is the smallest significance level at which I would reject $H_0$?

**6.** Suppose I reject $H_0$ at significance level 0.01. What can be said about the P-value of the data used?

**7.** Suppose $H_0$ is as in Chapter 4. In each of the following, answer "yes," "no," or "can't tell from information given" to the question "Should I reject $H_0$?" Assume the significance level stays the same.

(a) Our initial data makes us reject $H_0$. We then add on more data, while keeping the sample mean the same.

(b) Our initial data makes us not reject $H_0$. We then add on more data, while keeping the sample mean the same.

(c) Our initial data makes us reject $H_0$. We then take away some of the data, while keeping the sample mean the same.

(d) Our initial data makes us not reject $H_0$. We then take away some of the data, while keeping the sample mean the same.

**8.** Suppose $x_1, x_2, \ldots, x_{100}$ is a random sample from a normal random variable $X$, with $\sum_i x_i = 200$. Test

$$H_0 : \mu = 2.3 \;\; H_a : \mu < 2.3$$

at significance level 0.1 by first getting a rejection region for $H_0$. Assume $X$ has standard deviation 3.

**9.** Suppose $X$ is a normal random variable with standard deviation 5 and unknown mean $\mu$. We wish to test

$$H_0 : \mu = 10, \quad H_a : \mu \neq 10.$$

Get a P-value for data whose mean $\overline{x}$ equals 9.5, if

(a) the sample size $n$ equals 9;

(b) the sample size $n$ equals 100.

**10.** Which of the following two sets of data (choose one of (a) or (b)) has a smaller P-value, for the null hypothesis of my not having vandalized the Emperor's palace?

(a) My fingerprints are found in the palace;

(b) My fingerprints and my driver's license are found in the palace.

**11.** I hypothesize that female frogs are more than one gram heavier than male frogs, on average. To try to demonstrate this, I measure 100 female frogs and get a sample mean of 6 grams.

Assume female frog mass is normally distributed with a standard deviation of 5 grams and the average mass of male frogs is 4 grams.

Test my hypothesis at a significance level of 0.001. Include a precise statement of $H_0$, the null hypothesis, and $H_a$, the alternative hypothesis, both in terms of a well-defined parameter, calculate a P-value and use it to make a decision.

**12.** "Barking Fools" brand of dog food advertises that it will make dogs faster, on average. Test this claim, with significance level 0.01, if the average dog runs 30 miles per hour and randomly chosen dogs eating "Barking Fools" are normally distributed, with a standard deviation of 10 miles per hour, while the average running speed of 100 dogs eating "Barking Fools" is 32 miles per hour.

**13.** Same as number 12, except significance level of 0.05.

**14.** Assume human speech volume has an average of 85 decibels. If the average decibel level of 100 people who meditate is 80 decibels, does that provide conclusive evidence, at significance level 0.01, that meditation makes people (on average) quieter? Assume that randomly chosen people who meditate are normally distributed with a standard deviation of 20 decibels.

Include a precise statement of $H_0$, the null hypothesis, and $H_a$, the alternative hypothesis, both in terms of a well-defined parameter, calculate a P-value and use it to make a decision.

**15.** Same as number 14, except we are testing "meditation changes human speech volume."

**16.** A lotion called SkinTemp advertises that its use will change people's skin temperature, on average. Test this advertisement, at significance level 0.01, if the average skin temperature of 100 people using SkinTemp is 97 degrees. Assume that the skin temperature of a randomly chosen person has a mean of 99 degrees and the skin temperature of people who use SkinTemp is normally distributed, with a standard deviation of 5 degrees.

Use a rejection region.

**17.** Same as number 16, except we are testing "its use will lower people's skin temperature, on average."

**18.** Same as number 17, except we are testing "its use will lower people's skin temperature by at least one degree, on average."

**19.** For the hypothesis test of Examples 2.2(c.), fill in the missing numbers in the following table for the P-values corresponding to different values of $T \equiv$ number of heads when flipping fair coin ten times. See the APPENDIX.

| value of $T$ | P-value |
|---|---|
| 0 | $\binom{10}{0}(\frac{1}{2})^{10} + \binom{10}{10}(\frac{1}{2})^{10} = 2\binom{10}{0}(\frac{1}{2})^{10} = 2(\frac{1}{2})^{10} \sim 0.002$ |
| 1 | $\binom{10}{0}(\frac{1}{2})^{10} + \binom{10}{1}(\frac{1}{2})^{10} + \binom{10}{9}(\frac{1}{2})^{10} + \binom{10}{10}(\frac{1}{2})^{10} = 2\left[\binom{10}{0}(\frac{1}{2})^{10} + \binom{10}{1}(\frac{1}{2})^{10}\right] = 22(\frac{1}{2})^{10} \sim 0.0215$ |
| 2 | $2\left[\binom{10}{0}(\frac{1}{2})^{10} + \binom{10}{1}(\frac{1}{2})^{10} + \binom{10}{2}(\frac{1}{2})^{10}\right] = 112(\frac{1}{2})^{10} \sim 0.109$ |
| 3 | |
| 4 | |
| 5 | |
| 6 | |
| 7 | |
| 8 | |
| 9 | $22(\frac{1}{2})^{10} \sim 0.0215$ |
| 10 | $2(\frac{1}{2})^{10} \sim 0.002$ |

**20.** Suppose we are sampling from a normal population with standard deviation 30, and get a sample mean of 10, as in [2, HW1]. As in [2, HW1], we wish to perform statistical inference on $\mu$, the population mean.

(a) Suppose the sample size is 9.

Test, at significance level 0.1,

$$(i)\ H_0 : \mu = 0 \quad H_a : \mu \neq 0$$

and

$$(ii)\ H_0 : \mu = 30 \quad H_a : \mu \neq 30.$$

Compare to [2, HW1(a)].

(b) Suppose the sample size is 900.
Test, at significance level 0.1,

$$(i)\ H_0 : \mu = 0 \quad H_a : \mu \neq 0,$$
$$(ii)\ H_0 : \mu = 11 \quad H_a : \mu \neq 11,$$

and

$$(iii)\ H_0 : \mu = 8 \quad H_a : \mu \neq 8.$$

Compare to [2, HW1(b)].

(c) Suppose the sample size is 9.
Test, at significance level 0.01,

$$(i)\ H_0 : \mu = 0 \quad H_a : \mu \neq 0$$

and

$$(ii)\ H_0 : \mu = 30 \quad H_a : \mu \neq 30.$$

Compare to [2, HW1(c)].

(d) Suppose the sample size is 900.
Test, at significance level 0.01,

$$(i)\ H_0 : \mu = 0 \quad H_a : \mu \neq 0$$

and

$$(ii)\ H_0 : \mu = 8 \quad H_a : \mu \neq 8.$$

Compare to [2, HW1(d)].

**21.** Suppose data produces the 95% confidence interval $(2, 5)$ for $\mu$. Test each of the following, at significance level 0.05.

$$(i)\ H_0 : \mu = 0 \quad H_a : \mu \neq 0$$
$$(ii)\ H_0 : \mu = 3 \quad H_a : \mu \neq 3$$
$$(iii)\ H_0 : \mu = 6 \quad H_a : \mu \neq 6$$

**22.** Suppose data produces 0 as a 99% lower confidence bound for $\mu$. Test each of the following, at significance level 0.01.

$$(i)\ H_0 : \mu = 1 \quad H_a : \mu > 1$$
$$(ii)\ H_0 : \mu = 2.3 \quad H_a : \mu > 2.3$$
$$(iii)\ H_0 : \mu = -2 \quad H_a : \mu > -2$$

**23.** Suppose data produces 100 as a 90% upper confidence bound for $\mu$. Test each of the following, at significance level 0.1.

$$(i)\ H_0 : \mu = 0 \quad H_a : \mu < 0$$
$$(ii)\ H_0 : \mu = 200 \quad H_a : \mu < 200$$
$$(iii)\ H_0 : \mu = 50 \quad H_a : \mu < 50$$

## HOMEWORK ANSWERS

**1.** $\mu_0$ and $p_0$ below are (known) numbers, while $\mu$ and $p$ are unknown parameters.

**1.6(a).**

$$H_0 : \mu = \mu_0 \quad H_a : \mu > \mu_0,$$

where $\mu_0$ is the average weight, in pounds, of all American wolverines, and $\mu$ is the average weight, in pounds, of all Canadian wolverines.

**1.6(b).**

$$H_0 : \mu = \mu_0 \quad H_a : \mu > \mu_0,$$

where $\mu_0$ is the average annual income, in dollars, of all people, while $\mu$ is the average annual income, in dollars, of all people who sleep on the floor.

**1.6(c).**

$$H_0 : p = p_0 \quad H_a : p < p_0,$$

where $p_0$ is the proportion of people who get glaucoma, $p$ is the proportion of people exposed to colored light who get glaucoma.

**2.** (a) This is (ii), with $\mu_0$ the average weight of all people, $\mu$ the average weight of all vaccinated people.

(b) Same as (a), except (iii) instead of (ii).

(c) This is (i), with $\mu_0$ the average number of pounds a Freedonian can bench press, $\mu$ the average number of pounds a Brutopian can bench press.

There are many other correct interpretations of (c).

**3.** (b)

**4.** See Theorem 3.6 and the assumption at the beginning of the homework.

(a) yes, since P-value $\leq 0.01 \leq 0.05$ implies P-value $\leq 0.05$.

(b) no; we only know that P-value $\leq 0.01$, we can't tell if said P-value is $\leq 0.001$.

**5.** See Theorem 3.6 and the assumption at the beginning of the homework.

(a) no, since P-value is greater than 0.01.

(b) yes, since P-value is less than or equal to 0.05.

(c) 0.035.

**6.** P-value is less than or equal to 0.01; see Theorem 3.6 and the assumption at the beginning of the homework.

**7.** (a) yes, since the P-value (see 4.2) will get smaller, hence remain less than or equal to the significance level.

(b) can't tell, since the P-value will get smaller, but we can't tell if it gets less than or equal to the significance level.

(c) can't tell, since the P-value will get larger, but we can't tell if it gets larger than the significance level.

(d) no, since the P-value gets larger, hence remains larger than the significance level.

**8.** $\overline{x} = 2, \sigma = 3, n = 100$, and $\mu_0 = 2.3$, thus our test statistic is

$$z = \frac{(2 - 2.3)}{\frac{3}{\sqrt{100}}} = -1.$$

Since $\alpha = 0.1$, our critical value is $z_\alpha = z_{0.1} = 1.28$, thus our rejection region for $H_0$ is

$$z \leq -1.28;$$

since

$$z = -1 > -1.28 = -z_\alpha,$$

we do not reject $H_0$.

**9.** $\sigma = 5, \overline{x} = 9.5$, and $\mu_0 = 10$.

(a) $z = \frac{(9.5-10)}{\frac{5}{\sqrt{9}}} = -0.3$, so our P-value is

$$P(|Z| > 0.3) = 2P(Z > 0.3) = 0.7642.$$

(b) $z = \frac{(9.5-10)}{\frac{5}{\sqrt{100}}} = -1$, so our P-value is

$$P(|Z| > 1) = 2P(Z > 1) = 0.3174.$$

**10.** (b)

**11.** Let $\mu$ be the average mass of female frogs, in grams. Our hypothesis test is

$$H_0 : \mu = 5 \quad H_a : \mu > 5,$$

with $\sigma = 5, n = 100$, and $\overline{x} = 6$, so that our test statistic is

$$z = \frac{(6-5)}{\frac{5}{\sqrt{100}}} = 2,$$

and our P-value is

$$P(Z > 2) = 0.0228 > 0.001 = \alpha,$$

so we do not reject $H_0$; the data is insufficient, at significance level 0.001, to assert that female frogs are at least one gram heavier than male frogs.

**12.** Let $\mu$ be the average speed of Barking Fools (BF) consuming dogs, in miles per hour. Our hypothesis test is

$$H_0 : \mu = 30 \quad H_a : \mu > 30.$$

We have $\overline{x} = 32, n = 100$, and $\sigma = 10$, so that our test statistic is

$$z = \frac{(32-30)}{\frac{10}{\sqrt{100}}} = 2,$$

and our P-value is

$$P(Z > 2) = 0.0228 > 0.01 \equiv \alpha,$$

thus we do not reject $H_0$; our data does not provide 0.01 significance level evidence that BF makes dogs faster.

**13.** Now we have

$$\text{P-value} = 0.0228 \leq 0.05 \equiv \alpha,$$

so we reject $H_0$; our data does provide 0.05 significance level evidence that BF makes dogs faster.

**14.** Let $\mu$ be the average decibel level of people who meditate. Our hypothesis test is

$$H_0 : \mu = 85 \quad H_a : \mu < 85,$$

with $\sigma = 20, \overline{x} = 80$, and $n = 100$, so that

$$z = \frac{(80-85)}{\frac{20}{\sqrt{100}}} = -2.5,$$

and our P-value is

$$P(Z < -2.5) = P(Z > 2.5) = 0.0062 \leq 0.01 \equiv \alpha,$$

thus we reject $H_0$; there is sufficient evidence to conclude, at significance level 0.01, that meditation makes people quieter.

**15.** In the language of no. 14, we have

$$H_0 : \mu = 85 \quad H_a : \mu \neq 85,$$

which changes our P-value to

$$P(|Z| > 2.5) = 2P(Z > 2.5) = 0.0124 > 0.01,$$

so we do not reject $H_0$; there is insufficient evidence to conclude, at significance level 0.01, that meditation changes human speech volume.

See (i.) vs (j.), in More Examples 4.9.

**16.** Let $\mu$ be the average skin temperature of people using SkinTemp (ST). We have

$$H_0 : \mu = 99 \quad H_a : \mu \neq 99,$$

with $\overline{x} = 97, \sigma = 5$, and $n = 100$, so that

$$z = \frac{(97 - 99)}{\frac{5}{\sqrt{100}}} = -4.$$

Since $\alpha \equiv 0.01$, our rejection region is

$$|z| \geq z_{0.005} = 2.58.$$

Since $|z| = 4 \geq 2.58$, we reject $H_0$; the data suggests, at significance level 0.01, that ST changes skin temperature.

**17.** Our only change from no. 16 is

$$H_a : \mu < 99;$$

this changes the rejection region to

$$z \leq -z_{0.01} = -2.33.$$

Since our test statistic is $z = -4 \leq -2.33$, we reject $H_0$; the data suggests, at significance level 0.01, that ST lowers skin temperature.

**18.** Our hypothesis test is now

$$H_0 : \mu = 98 \quad H_a : \mu < 98,$$

thus (see no. 16) we now have test statistic

$$z = \frac{(97 - 98)}{\frac{5}{\sqrt{100}}} = -2.$$

Our rejection region is still, as in no. 17,

$$z \leq -z_{0.01} = -2.33.$$

Since our test statistic $z = -2 > -2.33 = -z_{0.01}$, we do not reject $H_0$; the data does not suggest, at significance level 0.01, that ST lowers skin temperature by at least one degree.

**19.**

| value of $T$ | P-value |
|---|---|
| 0 | $\binom{10}{0}(\frac{1}{2})^{10} + \binom{10}{10}(\frac{1}{2})^{10} = 2\binom{10}{0}(\frac{1}{2})^{10} = 2(\frac{1}{2})^{10} \sim 0.002$ |
| 1 | $\binom{10}{0}(\frac{1}{2})^{10} + \binom{10}{1}(\frac{1}{2})^{10} + \binom{10}{9}(\frac{1}{2})^{10} + \binom{10}{10}(\frac{1}{2})^{10} = 2\left[\binom{10}{0}(\frac{1}{2})^{10} + \binom{10}{1}(\frac{1}{2})^{10}\right] = 22(\frac{1}{2})^{10} \sim 0.0215$ |
| 2 | $2\left[\binom{10}{0}(\frac{1}{2})^{10} + \binom{10}{1}(\frac{1}{2})^{10} + \binom{10}{2}(\frac{1}{2})^{10}\right] = 112(\frac{1}{2})^{10} \sim 0.109$ |
| 3 | $2\left[\binom{10}{0}(\frac{1}{2})^{10} + \binom{10}{1}(\frac{1}{2})^{10} + \binom{10}{2}(\frac{1}{2})^{10} + \binom{10}{3}(\frac{1}{2})^{10}\right] = 352(\frac{1}{2})^{10} \sim 0.344$ |
| 4 | $2\left[\binom{10}{0}(\frac{1}{2})^{10} + \binom{10}{1}(\frac{1}{2})^{10} + \binom{10}{2}(\frac{1}{2})^{10} + \binom{10}{3}(\frac{1}{2})^{10} + \binom{10}{4}(\frac{1}{2})^{10}\right] = 772(\frac{1}{2})^{10} \sim 0.754$ |
| 5 | 1 |
| 6 | $772(\frac{1}{2})^{10} \sim 0.754$ |
| 7 | $352(\frac{1}{2})^{10} \sim 0.344$ |
| 8 | $112(\frac{1}{2})^{10} \sim 0.109$ |
| 9 | $22(\frac{1}{2})^{10} \sim 0.0215$ |
| 10 | $2(\frac{1}{2})^{10} \sim 0.002$ |

**20.** We'll do these with rejection regions (see 4.5); $\sigma = 30, \overline{x} = 10$.

(a) $n = 9, \alpha = 0.1$.

(i) $\mu_0 = 0$, so

$$|z| = |\frac{(10-0)}{\frac{30}{\sqrt{9}}}| = 1 < 1.65 = z_{\frac{\alpha}{2}},$$

so we don't reject $H_0$.

(i) $\mu_0 = 30$, so

$$|z| = |\frac{(10-30)}{\frac{30}{\sqrt{9}}}| = 2 \geq 1.65 = z_{\frac{\alpha}{2}},$$

so we reject $H_0$.

In [2, HW1(a)], we got a $100(1-0.1)\%$ confidence interval (CI) for $\mu$ of

$$(-6.5, 26.5) \equiv \{c \,|\, -6.5 < c < 26.5\};$$

0 is in the CI and 30 is not in the CI.

(b) $n = 900, \alpha = 0.1$.

(i) $\mu_0 = 0$, so

$$|z| = |\frac{(10-0)}{\frac{30}{\sqrt{900}}}| = 10 \geq 1.65 = z_{\frac{\alpha}{2}},$$

so we reject $H_0$.

(ii) $\mu_0 = 11$, so

$$|z| = |\frac{(10-11)}{\frac{30}{\sqrt{900}}}| = 1 < 1.65 = z_{\frac{\alpha}{2}},$$

so we don't reject $H_0$.

(iii) $\mu_0 = 8$, so

$$|z| = |\frac{(10-8)}{\frac{30}{\sqrt{900}}}| = 2 \geq 1.65 = z_{\frac{\alpha}{2}},$$

so we reject $H_0$.

In [2, HW1(b)], we got a $100(1-0.1)\%$ confidence interval (CI) for $\mu$ of

$$(8.35, 11.65) \equiv \{c \,|\, 8.35 < c < 11.65\};$$

11 is in the CI and 0 and 8 are not in the CI.

(c) This is the same as (a), except $\alpha = 0.01$.

(i) $\mu_0 = 0$: As in (a), $|z| = 1$, but now $z_{\frac{\alpha}{2}} = 2.58$, thus $|z| < z_{\frac{\alpha}{2}}$ and we don't reject $H_0$.

(ii) $\mu_0 = 30$: As in (a), $|z| = 2 < 2.58 = z_{\frac{\alpha}{2}}$, so we don't reject $H_0$.

In [2, HW1(c)], we got a $100(1-0.01)\%$ confidence interval (CI) for $\mu$ of

$$(-15.8, 35.8) \equiv \{c \,|\, -15.8 < c < 35.8\};$$

both 0 and 30 are in the CI.

(d) This is the same as (b)(i) and (iii), except $\alpha = 0.01$.

(i) $\mu_0 = 0$: As in (b), $|z| = 10$, but now $z_{\frac{\alpha}{2}} = 2.58$, thus $|z| \geq z_{\frac{\alpha}{2}}$ and we reject $H_0$.

(ii) $\mu_0 = 8$: As in (b), $|z| = 2 < 2.58 = z_{\frac{\alpha}{2}}$, so we don't reject $H_0$.

In [2, HW1(d)], we got a $100(1-0.01)\%$ confidence interval (CI) for $\mu$ of

$$(7.42, 12.58) \equiv \{c \,|\, 7.42 < c < 12.58\};$$

8 is in the CI and 0 is not in the CI.

**21.** Use Theorem 5.2(c).

For (i) and (iii), reject $H_0$, since neither 0 nor 6 are in the confidence interval; for (ii), don't reject $H_0$, since 3 is in the confidence interval.

**22.** Use Theorem 5.2(a).

For (iii), reject $H_0$, since $-2$ is less than or equal to 0. For (i) and (ii), don't reject $H_0$, since both 1 and 2.3 are greater than 0.

**23.** Use Theorem 5.2(b).

For (ii), reject $H_0$, since 200 is greater than or equal to 100. For (i) and (iii), don't reject $H_0$, since both 50 and 0 are less than 100.

## REFERENCES

**1.** R. deLaubenfels, "Probability and Counting Magnification" https://teacherscholarinstitute.com/MathMagnificationsReadyToUse.html.

**2.** R. deLaubenfels, "Statistics: Confidence Intervals Magnification," https://teacherscholarinstitute.com/MathMagnificationsReadyToUse.html.

**3.** R. deLaubenfels, "Statistics Introduction Magnification," https://teacherscholarinstitute.com/MathMagnificationsReadyToUse.html.

**4.** J. L. Devore, "Probability and Statistics for Engineering and the Sciences," Brooks/Cole, eighth edition, 2012.

**5.** J. Saxon, "Algebra 1. An Incremental Development," Second Edition, Saxon Publishers, Inc., 1990.