Hypothesis Tests

Hypothesis Tests as Statistical Inference

Recall from chapter 5 that statistical inference is the use of a subset of a population (the sample) to draw conclusions about the entire population.

There are two main branches of statistical inference: **estimation**, which includes the topics of point estimation and confidence intervals that we studied in chapter 5, and **hypothesis tests**, which we will study in the coming days.

To introduce the ideas behind hypothesis testing, we will look at the same example we used to introduce estimation.

Hypothesis Testing: An Example

Determining the Effectiveness of a Medication

A pharmaceutical company is testing the effectiveness of a new cholesterol-lowering medication. Specifically, they want to know if the medication reduces low density lipoprotein (LDL) level in people with high LDL.

$\approx$

Hypothesis Testing: An Example

To do so, the company's scientists propose the following study. They will

- Recruit 10 subjects with high levels of LDL cholesterol.
- Make sure the subjects don't take any cholesterol medication for two weeks to ensure an accurate baseline measurement.
- Take an initial baseline reading.
- Take a follow-up LDL measurement after the subject has been 30 days on the test medication.

≈

Hypothesis Testing: An Example

Here are the resulting data:

Subject	Baseline	Follow-up	LDL Decrease
1	160.5	168.1	−7.6
2	195.3	181.4	13.9
3	181.7	154.6	27.1
4	175.1	160.3	14.8
5	198.3	192.0	6.3
6	215.5	173.5	42.0
7	227.9	186.2	41.7
8	201.7	183.2	18.5
9	161.5	130.3	31.2
10	189.0	165.0	24.0

The Components of a Statistical Hypothesis Testing Problem

As we did for estimation problems, we will divide a statistical hypothesis testing problem into five steps

1. **The Scientific Hypothesis**
2. **The Statistical Model**
3. **The Statistical Hypotheses**
4. **The Test Statistic**
5. **The P-Value**

$\approx$

The Components of a Statistical Hypothesis Testing Problem

1. **The Scientific Hypothesis** The scientific hypothesis is the hypothesized outcome of the experiment or study. In this example, we will take the scientific hypothesis to be that, on average, the medication does lower LDL levels of people with high LDL.
2. **The Statistical Model** The statistical model is the distribution of the population of measurements that are being taken. In this case, the measurements are the LDL decreases and we will assume the population has a $N(\mu, \sigma^2)$ distribution.

The Components of a Statistical Hypothesis Testing Problem

3. **The Statistical Hypotheses** In hypothesis testing problems, we always state two competing hypotheses: H_0, the **null hypothesis** and H_a, the **alternative hypothesis**. For this example, we will take these to be

$$\begin{aligned} H_0: \quad \mu &= 0 \\ H_a: \quad \mu &> 0 \end{aligned}$$

Notice that H_a states the scientific hypothesis.

$\approx$

The Components of a Statistical Hypothesis Testing Problem

4. **The Test Statistic** In all one-parameter hypothesis test settings we will consider, the test statistic will be the estimator of the population parameter about which inference is being made. As you know from chapter 5, the estimator of μ is the sample mean, $\overline{y}$, and this is also the test statistic. The observed value of $\overline{y}$ for these data is $\overline{y}^* = 21.19$.

$\approx$

The Components of a Statistical Hypothesis Testing Problem

5. **The P-Value** Think of this as the **plausibility value**. To compute the p-value, we first assume H_0 is true. The p-value is then the proportion of all samples from this population for which the test statistic will give as much or more evidence against H_0 and in favor of H_a as does the observed test statistic value.

   That is, if H_0 is true, how plausible is the observed value of the test statistic?

The Components of a Statistical Hypothesis Testing Problem

For the LDL experiment, $H_0: \ \mu = 0$. Since H_a states that $\mu > 0$, large values of $\overline{y}$ will provide evidence against H_0 and in favor of H_a. Therefore, any value of $\overline{y}$ as large or larger than the observed value $\overline{y}^* = 21.19$ will provide as much or more evidence against H_0 and in favor of H_a as does the observed test statistic value.

Thus, the *p*-value is $Pr_0(\overline{y} \geq 21.19)$, where $Pr_0($*expression*$)$ represents the proportion of all samples for which *expression* is true, computed under the assumption that H_0 is true.

So $Pr_0(\overline{y} \geq 21.19)$ represents the proportion of all samples of size 10 from a $N(0, \sigma^2)$ population for which $\overline{y} \geq 21.19$.

$\approx$

The Components of a Statistical Hypothesis Testing Problem

To make things clear, let's first assume we know the population vstandard deviation σ. Here, we'll assume it equals 16.

To calculate the *p*-value, we standardize the test statistic by subtracting its mean (remember we're assuming H_0 is true, so this means we take $\mu = 0$) and dividing by its standard error, $\sigma/\sqrt{n}$:

$$(\overline{y} - 0)/(\sigma/\sqrt{n})$$

If H_0 is true, the result will have a ??? distribution.
Answer: $N(0, 1)$

$\approx$

The Components of a Statistical Hypothesis Testing Problem

For these data, $\sigma/\sqrt{n} = 16/\sqrt{10} = 5.060$, so the p-value is

$$\begin{aligned} Pr_0(\overline{y} \geq 21.19) &= Pr_0\left(\frac{\overline{y}-0}{\sigma/\sqrt{10}} \geq \frac{21.19-0}{\sigma/\sqrt{10}}\right) \\ &= Pr_0\left(\frac{\overline{y}-0}{5.060} \geq \frac{21.19-0}{5.060}\right) \\ &= Pr(N(0,1) \geq 4.188) \\ &= 1.4 \times 10^{-5} \end{aligned}$$

Computers, calculators and applets such as this can be used to find p-values.

The following picture might put this in perspective:

$\approx$

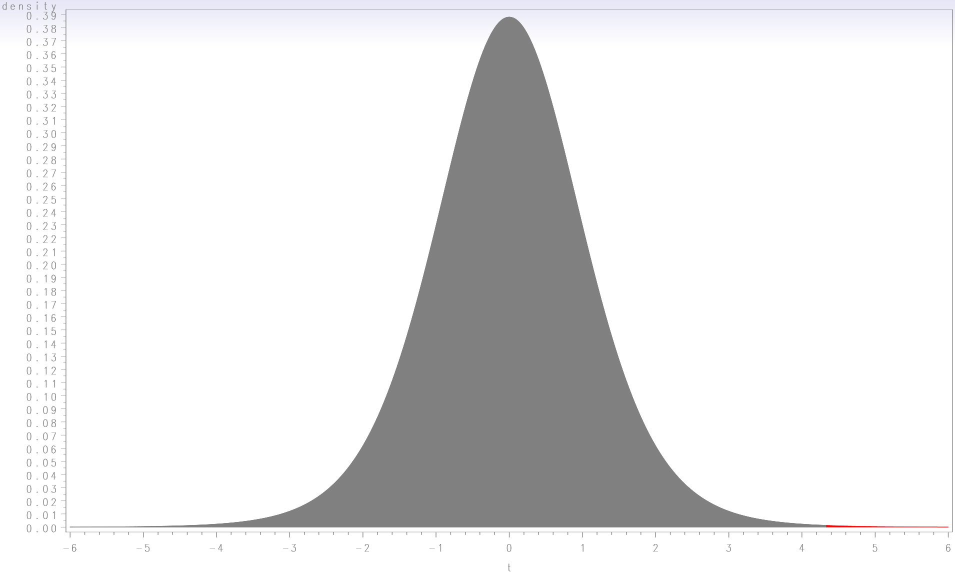

Figure: *1: The value of the standardized test statistic is 4.188. The p-value is the area in red to the right of this.*

The Components of a Statistical Hypothesis Testing Problem

If we don't know the population variance σ^2, we estimate it using the sample variance s^2. To calculate the p-value, we standardize the test statistic by subtracting its mean (remember we're assuming H_0 is true, so we take $\mu = 0$) and dividing by its estimated standard error, $s/\sqrt{n}$:

$$(\overline{y} - 0)/(s/\sqrt{n})$$

If H_0 is true, the result will have a ??? distribution.
Answer: $t_{n-1} = t_9$

$\approx$

The Components of a Statistical Hypothesis Testing Problem

It can be shown that the p-value is then

$$Pr(t_{n-1} \geq t^*),$$

where t^* is the observed value of the standardized test statistic:

$$(\overline{y} - 0)/(s/\sqrt{n}).$$

$\approx$

The Components of a Statistical Hypothesis Testing Problem

For these data, $s = 15.45$ and thus

$$t^* = \frac{21.19 - 0}{15.45/\sqrt{10}} = 4.337$$

The p-value (which can be computed with this applet) is

$$Pr(t_{n-1} \geq t^*) = Pr(t_9 \geq 4.337) = 9.4 \times 10^{-4}.$$

$\approx$

The Components of a Statistical Hypothesis Testing Problem

What's the Conclusion?

At this point, we have two options:

- Reject H_0 in favor of H_a.
- Do not reject H_0 in favor of H_a.

$\approx$

The Components of a Statistical Hypothesis Testing Problem

If the p-value is small enough, it indicates that, relative to H_a, the data are not consistent with the assumption that H_0 is true (that is, they are not **plausible**), so our action would be to reject H_0 in favor of H_a. (After all, if theory and data disagree, the data win!)

How small is "small enough" to reject H_0 in favor of H_a? That depends on a number of factors, such as the type of study, the purposes of the study, and the number of hypothesis tests being conducted. Table 1 gives guidelines for a single hypothesis test.

$\approx$

The Components of a Statistical Hypothesis Testing Problem

If the p-value is less than:	*The evidence against H_0 and in favor of H_a is:*
0.100	borderline
0.050	reasonably strong
0.025	strong
0.010	very strong

Table: 1: Interpreting the strength of evidence against H_0 and in favor of H_a provided by p-values

So, based on the p-value of 9.4×10^{-4}, the evidence against H_0 and in favor of H_a in the LDL experiment is very strong indeed.

We would undoubtedly reject H_0 in favor of H_a, and conclude that there is a positive mean decrease in LDL levels for the population.

Something About Hypotheses

In all examples we'll look at, H_0 will be **simple** (i.e. will state that the parameter has a single value) as opposed to **composite**. Alternative hypotheses will be **one-sided** (that the parameter be larger the null value, or smaller than the null value) or **two-sided** (that the parameter not equal the null value).[1]

In the LDL reduction example, we had

$$\begin{array}{lcll} H_0: & \mu & = & 0 \text{ (simple)} \\ H_a: & \mu & > & 0 \text{ (composite, one-sided)} \end{array}$$

$\approx$

---

[1] It turns out that the tests we will consider of simple versus a one-sided alternative hypothesis are also optimal for comparing two one-sided hypotheses, such as $H_0 : \mu \leq 0$ versus $H_a : \mu > 0$, but development of such tests is a topic for more advanced courses.

Two-Sided Tests

Suppose in the LDL reduction problem that the researchers wanted to see if there was no mean change in LDL levels as opposed to the medication making some difference, either good (a positive mean reduction) or bad (a negative mean reduction).

Then appropriate hypotheses would be:

$$\begin{array}{lcll} H_0: & \mu & = & 0 \text{ (simple)} \\ H_a: & \mu & \neq & 0 \text{ (composite, two-sided)} \end{array}$$

In this case, evidence against H_0 and in favor of H_a is provided by both large and small values of $\overline{y}$.

$\approx$

Two-Sided Tests

To compute the p-value of the two-sided test, we first compute the standardized test statistic t, and its observed value, t^*:

$$t = \frac{\overline{y} - \mu_0}{S/\sqrt{n}}, \; t^* = \frac{21.19 - 0}{4.886} = 4.337.$$

Recall that under H_0, $t \sim t_9$.

$\approx$

Two-Sided Tests

Because the test is two-sided, we compute the *p*-value as $Pr(|t_9| \geq |t^*|) = Pr(t_9 \leq -|t^*|) + Pr(t_9 \geq |t^*|)$. By the symmetry of the t_9 distribution about 0, this equals $2Pr(t_9 \geq |t^*|)$. For the present example, the p-value is $Pr(|t_9| \geq 4.337) = 2Pr(t_9 \geq 4.337) = 18.8 \times 10^{-4}$.

In what follows, the following formula for the two-sided p-value will prove useful:

$$2Pr(t_9 \geq |t^*|) = 2\min(Pr(t_9 \leq t^*), Pr(t_9 \geq t^*)).$$

This gives for the LDL example
$2\min(Pr(t_9 \leq 4.337), Pr(t_9 \geq 4.337))$

$$= 2\min(0.99906, 9.4 \times 10^{-4}) = 18.8 \times 10^{-4}.$$

$\approx$

Testing A Population Mean: Summary

Assumptions:

A random sample of n observations from a $N(\mu, \sigma^2)$ population (or n is large enough for the CLT).

Observed Value of Standardized Test Statistic:

$$t^* = \frac{\overline{y}^* - \mu_0}{s/\sqrt{n}}.$$

P-values:

H_0	H_a	p-value
$\mu = \mu_0$	$\mu < \mu_0$	$p_- = Pr(t_{n-1} \leq t^*)$
$\mu = \mu_0$	$\mu > \mu_0$	$p_+ = Pr(t_{n-1} \geq t^*)$
$\mu = \mu_0$	$\mu \neq \mu_0$	$p\pm = 2\min(p_-, p^+)$

$\approx$

The Philosophy of Hypothesis Testing

Statistical hypothesis testing is modeled on scientific investigation. The two hypotheses represent competing scientific hypotheses.

- The **alternative hypothesis** is the hypothesis that suggests change, difference or an aspect of a new theory.
- The **null hypothesis** is the hypothesis that represents the accepted scientific view. Most often, it is the hypothesis that suggests no difference or effect.

≈

The Philosophy of Hypothesis Testing

Example:

Quality testing of parts manufactured by an injection molding process includes a compressive strength test. Historically, the mean compressive strength has been 82.15 psi. Manufacturing personnel want to test whether the mean for present production matches the historical mean. Tests on a random sample of 25 parts taken from production yield a mean of 79.65 psi and a standard deviation of 5.1.

For this problem,

$H_0 : \mu = 82.15$, where μ is the mean strength of present production.
$H_a : \mu \neq 82.15$.

$\approx$

The Philosophy of Hypothesis Testing

Example:

It is well-known that the leading blood pressure drug reduces systolic readings by an average of 16.8 mmHg in people with unmedicated readings over 175 mmHg. A new medication is undergoing tests. Researchers have solid evidence that mean reduction from the new drug is at least as large as from the current drug; the question is whether it is larger. A random sample of 12 individuals in the at-risk group taking the new medication showed mean reduction of 23.7 mmHg with a standard deviation of 26.6 mmHg.

For this problem,

$H_0 : \mu = 16.8$, where μ is the mean systolic BP reduction under the new medication.
$H_a : \mu > 16.8$.

$\approx$

The Philosophy of Hypothesis Testing

Because we do not want to lightly overturn the accepted scientific view, the null hypothesis is given favored treatment. Such treatment consists of assuming H_0 is true when computing the p-value, and requiring compelling evidence (in the form of a small p-value) before rejecting that assumption. (Remember that the p-value is the “plausibility value”, which measures how consistent the data are with the assumption that H_0 is true.)

$\approx$

The Philosophy of Hypothesis Testing

A good analogy is to the Anglo-American criminal justice system, in which the byword is "innocent until proven guilty".

In this analogy, H_0 is on trial, accused of pretending to be true when in reality H_a is.

- Throughout the trial H_0 is presumed to be "innocent" (i.e., true).
- A "conviction" results in H_0 being rejected in favor of H_a.
- The "jury" (i.e., the scientific researchers) must not convict unless there is sufficient evidence–"beyond a reasonable doubt"– against H_0 and in favor of H_a, as measured by the p-value.

The Philosophy of Hypothesis Testing

Note that

- We presume H_0 to be true ("innocent") until the evidence indicates otherwise.
- The truth ("innocence") of H_0 is a presumption–it does not have to be proven.
- Failing to "convict" does not prove H_0 is true ("innocent"); it only means there is not enough evidence to reject H_0 in favor of H_a.

$\approx$

Statistical Significance

Often, prior to conducting a study, users of hypothesis tests set a pre-specified threshold level of evidence against the null and in favor of the alternative hypothesis. The rule is that H_0 will be rejected in favor of H_a if and only if the p-value falls below this threshold. The name given to this threshold is “signifance level”, and it is often denoted α.

If, for example, we decide to use a significance level of $\alpha = 0.05$, our action would be to reject H_0 in favor of H_a if the p-value is less than 0.05, and to not reject otherwise.

The significance level **must** be set before the data are analyzed, and ideally at the study design stage. To do otherwise risks biasing the results.

Statistical Significance

Table 1, reproduced here, can serve as a guide to selecting an appropriate significance level:

If the p-value is less than:	*The evidence against H_0 and in favor of H_a is:*
0.100	borderline
0.050	reasonably strong
0.025	strong
0.010	very strong

Statistical Significance

Given a significance level, we can often use statistical tables to decide whether to reject or not reject the null hypothesis. An example of how to do this can be found here.

≈

Statistical vs. Practical Significance

Statistical significance measures our ability to detect a difference. As such, it is a reflection of not only the size of the difference but also the amount of data (i.e., evidence) we have.

For instance, recall the LDL reduction example. There were 10 subjects having mean LDL reduction 21.19 and standard deviation 15.45.

$\approx$

Statistical vs. Practical Significance

To test

$$\begin{aligned} H_0: \quad \mu &= 0 \\ H_a: \quad \mu &> 0 \end{aligned}$$

we computed the *p*-value as

$$Pr_0\left(\frac{\overline{y}-0}{4.886} \geq \frac{21.19-0}{4.886}\right)$$

$$= Pr(t_9 \geq 4.337) = Pr(t_9 \geq 4.337) = 9.4 \times 10^{-4}.$$

(Recall that 4.886 was the estimated standard error of the mean computed as $s/\sqrt{n} = 15.45/\sqrt{10} = 4.886$)

$\approx$

Statistical vs. Practical Significance

Now suppose that we had a sample of size 3 with the same mean and standard deviation. The observed difference is the same as previously: $\overline{y} - 0 = 21.19$.

However, the estimated standard error is much larger: $s/\sqrt{n} = 15.45/\sqrt{3} = 8.920$, and the corresponding p-value is

$$Pr_0\left(\frac{\overline{y}-0}{8.92} \geq \frac{21.19-0}{8.92}\right)$$

$$= Pr(t_2 \geq 2.376) = Pr(t_2 \geq 2.376) = 0.0703.$$

$\approx$

Statistical vs. Practical Significance

So with a sample of size 10, a mean LDL decrease of 21.19 is highly significant, while for a sample of size 3 it is at best borderline significant.

The practical significance of the result does not change. What changes is the statistical significance: our ability to detect the difference as being distinct from the background variation (or "noise").

Be careful about...

- Using the same data to suggest and test hypotheses (Exploratory vs. confirmatory studies).
- Doing lots of tests results in lots of false positives.
- Equating lack of significance with failure.

≈

Testing A Population Proportion

Suppose we have a population whose members can be observed to have or not have a certain characteristic. Let p be the proportion of population members who have the characteristic.

Hypotheses about p are frequently tested by taking a random sample and observing the number in the sample who have the characteristic, as the following example illustrates.

$\approx$

Testing A Population Proportion

Example: Researchers in the LDL decrease experiment were interested in testing whether the medication would result in reductions in LDL levels in more than half of all people with high LDL.

Specifically, if p is the proportion of that population for whom the medication would result in reductions in LDL levels, they wanted to test

$$\begin{aligned} H_0: \quad p &= 0.5 \\ H_a: \quad p &> 0.5 \end{aligned}$$

$\approx$

Testing A Population Proportion

Recall that for the estimator of p, we used the sample proportion, $\hat{p} = y/n$, where y is the number in the sample who have the characteristic, and n is the sample size.

The test statistic we will use for this problem is y, which gives the same results as we would obtain using $\hat{p}$, and which is easier to work with.

$\approx$

Testing A Population Proportion

Since H_a specifies that p is larger than 0.5, large values of y should lead to rejection of H_0 in favor of H_a. Therefore, the p-value will be given by $Pr_0(y \geq y^*)$, where y^* is the number in the sample who have the characteristic, and $Pr_0(y \geq y^*)$ is the proportion under H_0 (that is, when $p = 0.5$), of all size n samples from the population for which y will be at least as large as y^*.

For the LDL example, $y^* = 9$, so the p-value is $Pr_0(y \geq 9)$. In order to compute this, we need the sampling distribution of y.

Testing A Population Proportion

Consider taking simple random samples of size n from a large population in which a proportion p have a given characteristic. The proportion of all those random samples in which exactly k of the n have the characteristic is given by the formula

$$Pr(y = k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ for } k = 0, 1, 2, \ldots, n,$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

and ! denotes "factorial".

This is the sampling distribution of y. It is called the **binomial** distribution with parameters n and p (abbreviated $b(n, p)$).

Testing A Population Proportion

In the LDL example, $n = 10$, so when $p = 0.5$, the sampling distribution of y, the number of subjects with lower LDL, is $b(10, 0.5)$. Thus, when H_0 is true, the proportion of all samples of size 10 in which exactly k of the 10 have lower LDL is

$$Pr_0(y = k) = \binom{10}{k} 0.5^k (1 - 0.5)^{10-k}, \; k = 0, 1, 2, \ldots, 10.$$

$\approx$

Testing A Population Proportion

Recall that in the LDL study, the number in the sample with reduced LDL is $y^* = 9$, and the p-value of the test is

$$\begin{aligned}
Pr_0(y \geq 9) &= \sum_{k=9}^{10} Pr_0(y = k) \\
&= \sum_{k=9}^{10} \binom{10}{k} 0.5^k (1-0.5)^{10-k} \\
&= \binom{10}{9} 0.5^9 (1-0.5)^1 + \binom{10}{10} 0.5^{10} (1-0.5)^0 \\
&= (10) 0.5^9 (1-0.5)^1 + 0.5^{10} (1-0.5)^0 \\
&= (11) 0.5^{10} = 0.0107.
\end{aligned}$$

Thus, the data provide strong evidence against H_0 and in favor of H_a, and we might justifiably conclude that the medication would reduce the LDL levels of more than half the population.

Testing A Population Proportion

The usual resources are available for computing a p-value from the binomial distribution. This applet will do the job, as will many calculators and the table found in the text.

≈

Testing A Population Proportion: Summary

Assumptions:
A random sample of n observations from a large population with a proportion p having a characteristic of interest.

Observed Value of Test Statistic:
y^*, the number in the sample having the characteristic.

P-values:

H_0	H_a	p-value
$p = p_0$	$p < p_0$	$p_- = Pr(b(n, p_0) \leq y^*)$
$p = p_0$	$p > p_0$	$p_+ = Pr(b(n, p_0) \geq y^*)$
$p = p_0$	$p \neq p_0$	See explanation below

$\approx$

Testing A Population Proportion

Unlike the test for population mean, the p-value for a two-sided test of a proportion is not generally twice the smaller of the p-values of the one-sided tests. The reason is that unless $p_0 = 0.5$, the $b(n, p_0)$ distribution is not symmetric.

As a result, we compute the p-value for a two-sided test as follows.

$\approx$

Testing A Population Proportion

Let

$$f(y) = \binom{n}{y} p_0^y (1 - p_0)^{n-y},\ y = 0, 1, \ldots, n$$

define the sampling distribution of y when $p = p_0$, and let y^* denote the value of y observed in the sample. The p-value of the test of $H_0 : p = p_0$ versus $H_a : p \neq p_0$ is obtained by summing all the values of $f(y)$ that are less than or equal to $f(y^*)$.

The following example will illlustrate.

Testing A Population Proportion

Example
A manufacturer of high fiber cereal claims that its product, Sawdusties, is recommended by 2 out of 3 nutritionists. In a small (but well-conducted) survey, 3 of a random sample of 6 nutritionists recommended Sawdusties. We want to test whether their claim is true for the population of nutritionists.

If p is the proportion of all nutritionists who recommend Sawdusties, the statistical hypotheses are

$$\begin{aligned} H_0 : \quad p &= 2/3 \\ H_a : \quad p &\neq 2/3 \end{aligned}$$

Testing A Population Proportion

Under H_0, y, the number of a sample of 6 who recommend Sawdusties has a $b(6, 2/3)$ distribution, so

$$f(y) = \left(\begin{array}{c} 6 \\ y \end{array} \right) (2/3)^y (1 - 2/3)^{6-y}, \;\; y = 0, 1, \ldots, 6.$$

Evaluating, we find:

$\approx$

Testing A Population Proportion

y	$f(y)$	y	$f(y)$
0	0.001	4	0.329
1	0.017	5	0.264
2	0.082	6	0.088
3	0.219		

The observed value of y is $y^* = 3$. The p value is the sum of all $f(y)$ values that are less than or equal to $f(y^*) = f(3) = 0.219$: That is, the p value equals

$$f(0) + f(1) + f(2) + f(3) + f(6)$$

$$= 0.001 + 0.017 + 0.082 + 0.219 + 0.088 = 0.407.$$

$\approx$

Testing A Population Proportion

The test based on the binomial distribution is called an **exact test** because it is based on the exact sampling distribution of the test statistic y. This is the preferred test and should be accessible to anyone with a good statistics computer program. A SAS macro to do both one and two-sided exact tests can be found here.

With a little additional calculation, it is do-able with a good statistics calculator (the TI-83, for instance, though the test computed under the STAT menu is not this exact test, but rather....)

≈

Testing A Population Proportion: Large Samples

...an approximate test that can also be done with only a basic calculator and a normal probability table. The approximation will be good if y^* and $n - y^*$ are both at least 10.

The test relies on the fact that under these conditions, the Central Limit Theorem ensures that the sampling distribution of y is approximately $N(np, np(1-p))$.

$\approx$

Testing A Population Proportion: Large Samples

The test is easy to summarize:

Assumptions:

A random sample of n observations from a large population with a proportion p having a characteristic of interest; y^* and $n - y^*$ should both be large enough (at least 10).

Observed Value of Standardized Test Statistic (with Continuity Corrections in red):

$$z_l^* = \frac{y^* - np_0 + 0.5}{\sqrt{np_0(1-p_0)}}, \quad z_u^* = \frac{y^* - np_0 - 0.5}{\sqrt{np_0(1-p_0)}}.$$

P-values:

H_0	H_a	p-value
$p = p_0$	$p < p_0$	$p_- = Pr(N(0,1) \leq z_l^*)$
$p = p_0$	$p > p_0$	$p_+ = Pr(N(0,1) \geq z_u^*)$
$p = p_0$	$p \neq p_0$	$p\pm = 2\min(p_-, p^+)$

$\approx$

Testing A Population Proportion: Large Samples

Now, for illustration, we will use the large-sample test on the Sawdusties data.

Neither $y^* = 3$ nor $n - y^* = 3$ are anywhere near 10, so using this test is not justified, and we would not use this test in practice, so this is just for illustration:

$$z_l^* = \frac{y^* - np_0 + 0.5}{\sqrt{np_0(1-p_0)}} = \frac{3 - (6)(2/3) + 0.5}{\sqrt{6(2/3)(1-2/3)}} = -0.433,$$

$$z_u^* = \frac{y^* - np_0 - 0.5}{\sqrt{np_0(1-p_0)}} = \frac{3 - (6)(2/3) - 0.5}{\sqrt{6(2/3)(1-2/3)}} = -1.299,$$

so $p_- = Pr(N(0,1) \leq -0.433) = 0.333$, and $p_+ = Pr(N(0,1) \geq -1.299) = 0.903$, so $p\pm = 2\min(0.333, 0.903) = 2 \times 0.333 = 0.666$ (recall that the exact value is 0.407).

$\approx$

Comparing Two Population Means: Paired Data

Recall the LDL reduction study in which each of 10 subjects obtained in a simple random sample was measured for LDL at the outset and then after 30 days on a particular medication. Our analysis focused on testing the mean reduction in LDL.

But viewed another way, this mean reduction is the difference of two population means: $\mu_1 - \mu_2$, where μ_1 is the mean of the before (or untreated) population, and μ_2 is the mean of the after (or treated) population.

Comparing Two Population Means: Paired Data

The before and after measurements are said to be **paired**, because each individual provides one before-after pair. Pairing is an example of blocking, as you studied in Chapter 3.

When paired data measurements are of the same quantity, as in the example, analysis is often done by subtracting one paired measurement from the other and treating the resulting difference as if it were a single measurement.

This is exactly what we did by choosing to analyze the decrease in LDL.

Comparing Two Population Means: Paired Data

The result is that the hypothesis test we conducted using the LDL decrease could be interpreted as a test of the hypotheses:

$$\begin{aligned} H_0: \quad \mu_1 - \mu_2 &= 0 \\ H_a: \quad \mu_1 - \mu_2 &> 0. \end{aligned}$$

The test statistic and p-value are identical to those obtained previously.

$\approx$

Comparing Two Population Means: Paired Data : Summary

Assumptions: A random sample of n paired observations, $(y_{1,i}, y_{2,i})$, $i = 1, \ldots, n$, where

(a) The population means of the $y_{1,i}$ and $y_{2,i}$ are μ_1 and μ_2, and

(b) $d_i = y_{1,i} - y_{2,i}$ are a random sample from a $N(\mu_1 - \mu_2, \sigma^2)$ population (or n is large enough for the CLT).

Observed Value of Standardized Test Statistic:

$$t^* = \frac{\overline{d}^* - \delta_0}{s_d/\sqrt{n}}.$$

P-values:

H_0	H_a	p-value
$\mu_1 - \mu_2 = \delta_0$	$\mu_1 - \mu_2 < \delta_0$	$p_- = Pr(t_{n-1} \leq t^*)$
$\mu_1 - \mu_2 = \delta_0$	$\mu_1 - \mu_2 > \delta_0$	$p_+ = Pr(t_{n-1} \geq t^*)$
$\mu_1 - \mu_2 = \delta_0$	$\mu_1 - \mu_2 \neq \delta_0$	$p\pm = 2\min(p_-, p^+)$

Comparing Two Population Means: Independent Populations

Not all comparisons are done using paired data. Sometimes our data consist of independent random samples from two separate populations.

Suppose that we take a random sample of size n_1 from population 1, which follows a $N(\mu_1, \sigma^2)$ distribution, and independently a random sample of size n_2 from population 2, which follows a $N(\mu_2, \sigma^2)$ distribution.

Notice that the only possible difference in the population distributions is in their means.

Comparing Two Population Means: Independent Populations

We already know that the estimator of μ_1 is the sample mean of the first sample, $\overline{y}_1$, and that of μ_2 is the sample mean of the second sample, $\overline{y}_2$.

We also know that the sampling distribution of $\overline{y}_1$ is $N(\mu_1, \sigma^2/n_1)$ and that of $\overline{y}_2$ is $N(\mu_2, \sigma^2/n_2)$.

Comparing Two Population Means: Independent Populations

In Chapter 5, we used $\overline{y}_1 - \overline{y}_2$ to estimate $\mu_1 - \mu_2$.

We also learned that the sampling distribution of the standardized estimator

$$t^{(p)} = \frac{\overline{y}_1 - \overline{y}_2 - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

has a $t_{n_1+n_2-2}$ distribution, where s_p is the pooled variance estmate of σ:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

$\approx$

Comparing Two Population Means: Independent Populations

If the null hypothesis is $H_0 : \mu_1 - \mu_2 = \delta_0$, then when H_0 is true,

$$t^{(p)} = \frac{\overline{y}_1 - \overline{y}_2 - \delta_0}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

From this, assuming $t^{(p)*}$ is the observed value of $t^{(p)}$, computed under the assumption that H_0 is true, we get the following hypothesis tests:

H_0	H_a	p-value
$\mu_1 - \mu_2 = \delta_0$	$\mu_1 - \mu_2 < \delta_0$	$p_- = Pr(t_{n_1+n_2-2} \leq t^{(p)*})$
$\mu_1 - \mu_2 = \delta_0$	$\mu_1 - \mu_2 > \delta_0$	$p_+ = Pr(t_{n_1+n_2-2} \geq t^{(p)*})$
$\mu_1 - \mu_2 = \delta_0$	$\mu_1 - \mu_2 \neq \delta_0$	$p\pm = 2\min(p_-, p^+)$

$\approx$

Example

A company buys cutting blades used in its manufacturing process from two suppliers. In order to decide if there is a difference in blade life, the lifetimes of 10 blades from manufacturer 1 and 13 blades from manufacturer 2 used in the same application are compared. A summary of the data shows the following (units are hours):

Manufacturer	n	$\overline{y}$	s
1	10	108.4	26.9
2	13	134.9	18.4

The manufacturer wants to test the equality of mean blade lives at the 0.10 level of significance, and is willing to assume the two population variances are equal.

Comparing Two Population Means: Independent Populations

1. **The Scientific Hypothesis** There is a difference in the lifetimes of blades from the two manufacturers.
2. **The Statistical Model** The two independent normal populations with equal variances.
3. **The Statistical Hypotheses**

$$\begin{aligned} H_0: \quad \mu_1 - \mu_2 &= 0 \\ H_a: \quad \mu_1 - \mu_2 &\neq 0 \end{aligned}$$

(Notice that here $\delta_0 = 0$)

$\approx$

Comparing Two Population Means: Independent Populations

4. **The Test Statistic**

The pooled variance estimate is

$$\begin{aligned} s_p^2 &= \frac{(10-1)(26.9)^2 + (13-1)(18.4)^2}{10+13-2} \\ &= 503.6, \end{aligned}$$

So the standard error estimate of $\overline{y}_1 - \overline{y}_2$ is

$$\begin{aligned} \hat{\sigma}_p(\overline{y}_1 - \overline{y}_2) &= \sqrt{503.6\left(\frac{1}{10}+\frac{1}{13}\right)} \\ &= 9.44. \end{aligned}$$

Therefore, $t^{(p)*} = (108.4 - 134.9)/9.44 = -2.81$, with 10+13-2=21 degrees of freedom.

$\approx$

Comparing Two Population Means: Independent Populations

5. **The p-value**
   $p_{-} = P(t_{21} \leq -2.81) = 0.0052$,
   $p^{+} = P(t_{21} \geq -2.81) = 0.9948$, and the p-value for this problem is $2\min(0.0052, 0.9948) = 0.0104$.

   Since the p-value is less than 0.10, we reject H_0 in favor of H_a, and conclude there is a difference in mean blade lifetimes.

   A SAS macro to conduct two-sample t tests using summary data is found here, and SAS code to conduct them using raw data here.

$\approx$

Comparing Two Population Means: Independent Populations

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

As we stated in Chapter 5, the most fundamental question is: "If the population variances are not equal, does it make sense to compare the population means?"

In the case of unequal variances, even if the means are equal, the two populations will have different distributions. So comparing the means is inappropriate if the goal is to decide if the two population distributions are the same.

$\approx$

Comparing Two Population Means: Independent Populations

If our interest is solely in comparing the means, and we are not willing to assume the population variances are equal, we can make use of the fact that the sampling distribution of the standardized estimator

$$t^{(ap)} = \frac{\overline{y}_1 - \overline{y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

can be approximated by a t distribution with ν degrees of freedom, where ν is the largest integer less than or equal to

$$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

$\approx$

Comparing Two Population Means: Independent Populations

From this, assuming $t^{(ap)*}$ is the observed value of $t^{(ap)}$, computed under the assumption that H_0 is true,

$$t^{(ap)*} = \frac{\overline{y}_1 - \overline{y}_2 - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

we get the following hypothesis tests:

H_0	H_a	p-value
$\mu_1 - \mu_2 = \delta_0$	$\mu_1 - \mu_2 < \delta_0$	$p_- = Pr(t_\nu \leq t^{(ap)*})$
$\mu_1 - \mu_2 = \delta_0$	$\mu_1 - \mu_2 > \delta_0$	$p_+ = Pr(t_\nu \geq t^{(ap)*})$
$\mu_1 - \mu_2 = \delta_0$	$\mu_1 - \mu_2 \neq \delta_0$	$p\pm = 2\min(p_-, p^+)$

$\approx$

Example

Recall the previous example: a company buys cutting blades used in its manufacturing process from two suppliers. In order to decide if there is a difference in blade life, the lifetimes of 10 blades from manufacturer 1 and 13 blades from manufacturer 2 used in the same application are compared. A summary of the data shows the following (units are hours):

Manufacturer	n	$\bar{y}$	s
1	10	108.4	26.9
2	13	134.9	18.4

The manufacturer wants to test the equality of mean blade lives at the 0.10 level of significance, but this time is unwilling to assume the population variances are equal.

$\approx$

Comparing Two Population Means: Independent Populations

1. **The Scientific Hypothesis** There is a difference in the lifetimes of blades from the two manufacturers.
2. **The Statistical Model** The two independent normal populations.
3. **The Statistical Hypotheses**

$$\begin{aligned} H_0: \quad \mu_1 - \mu_2 &= 0 \\ H_a: \quad \mu_1 - \mu_2 &\neq 0 \end{aligned}$$

$\approx$

Comparing Two Population Means: Independent Populations

4. **The Test Statistic**
   The observed value of the test statistic is

$$t^{(ap)*} = \frac{108.4 - 134.9}{\sqrt{\frac{(26.9)^2}{10} + \frac{(18.4)^2}{13}}} = -2.67.$$

$\approx$

Comparing Two Population Means: Independent Populations

5. **The p-value**

   The degrees of freedom ν is computed as the greatest integer less than or equal to

$$\frac{\left(\frac{(26.9)^2}{10}+\frac{(18.4)^2}{13}\right)^2}{\frac{\left(\frac{(26.9)^2}{10}\right)^2}{10-1}+\frac{\left(\frac{(18.4)^2}{13}\right)^2}{13-1}} = 15.17,$$

   so $\nu = 15$.
   $p_- = P(t_{15} \leq -2.67) = 0.0087$,
   $p^+ = P(t_{15} \geq -2.67) = 0.9913$, and the p-value for this problem is $2\min(0.0087, 0.9913) = 0.0174$.

$\approx$

Comparing Two Population Means: Independent Populations

Since the p-value is less than 0.10, we reject H_0 in favor of H_a, and conclude there is a difference in mean blade lifetimes.

The p-value in this case, 0.0174, is similar to the p-value we obtained in the pooled variance case: 0.0104, and the conclusion is the same.

A SAS macro to conduct two-sample t tests using summary data is found here, and SAS code to conduct them using raw data here.

$\approx$

Comparing Two Population Proportions

Suppose there are two populations: population 1, in which a proportion p_1 have a certain characteristic, and population 2, in which a proportion p_2 have a certain (possibly different) characteristic. We will use a sample of size n_1 from population 1, and n_2 from population 2 to test hypotheses about the difference $p_1 - p_2$.

Comparing Two Population Proportions

Specifically, if y_1 is the number having the population 1 characteristic in the n_1 items in sample 1, and if y_2 is the number having the population 2 characteristic in the n_2 items in sample 2, then the sample proportion having the population 1 characteristic is $\hat{p}_1 = y_1/n_1$, and the sample proportion having the population 2 characteristic is $\hat{p}_2 = y_2/n_2$.

A point estimator of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$. We will use this difference as our test statistic.

$\approx$

Comparing Two Population Proportions

The standard error of $\hat{p}_1 - \hat{p}_2$ is

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

Further, for large n_1 and n_2, the Central Limit Theorem ensures that $\hat{p}_1 - \hat{p}_2$ has approximately a normal distribution, so

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

has approximately a $N(0,1)$ distribution.

Comparing Two Population Proportions

We wish to test a null hypothesis that the two population proportions differ by a known amount δ_0,

$$H_0: \quad p_1 - p_2 \quad = \quad \delta_0,$$

against one of three possible alternative hypotheses:

$$\begin{aligned} H_{a^+}: &\quad p_1 - p_2 \quad > \quad \delta_0 \\ H_{a_-}: &\quad p_1 - p_2 \quad < \quad \delta_0 \\ H_{a\pm}: &\quad p_1 - p_2 \quad \neq \quad \delta_0 \end{aligned}$$

$\approx$

Comparing Two Population Proportions

The tests we will present rely on the normal approximation promised by the Central Limit Theorem. Therefore, you should always check that the sample sizes are large enough to justify this approximation. $y_i \geq 10$ and $n_i - y_i \geq 10$, $i = 1, 2$, suffices as a rule of thumb.

$\approx$

Comparing Two Population Proportions

Case 1: $\delta_0 = 0$

Suppose H_0 is $p_1 - p_2 = 0$. Then, let $p = p_1 = p_2$ denote the common value of the two population proportions. If H_0 is true, the variance of $\hat{p}_1$ equals $p(1-p)/n_1$ and that of $\hat{p}_2$ equals $p(1-p)/n_2$. This implies the standard error of $\hat{p}_1 - \hat{p}_2$ equals

$$\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Since we don't know p, we estimate it using the data from both populations:

$$\hat{p} = \frac{y_1 + y_2}{n_1 + n_2}.$$

$\approx$

Comparing Two Population Proportions

The estimated standard error of $\hat{p}_1 - \hat{p}_2$ is then

$$\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)},$$

and the standardized test statistic is then

$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},$$

which has approximately a $N(0,1)$ distribution if H_0 is true.

The observed value of Z_0 is denoted z_0^*.

Comparing Two Population Proportions when $\delta_0 = 0$

Assumptions:

A random sample of n_i observations from population i with a proportion p_i having a characteristic of interest, and $y_i \geq 10$, $n_i - y_i \geq 10$, $i = 1, 2$.

Observed Value of Standardized Test Statistic:

$$z_0^* = (\hat{p}_1 - \hat{p}_2) \Big/ \sqrt{\hat{p}(1-\hat{p})\,(1/n_1 + 1/n_2)}$$

Approximate P-values:

H_0	H_a	p-value
$p_1 - p_2 = 0$	$p_1 - p_2 < 0$	$p_- = Pr(N(0,1) \leq z_0^*)$
$p_1 - p_2 = 0$	$p_1 - p_2 > 0$	$p^+ = Pr(N(0,1) \geq z_0^*)$
$p_1 - p_2 = 0$	$p_1 - p_2 \neq 0$	$p\pm = 2\min(p_-, p^+)$

$\approx$

Comparing Two Population Proportions

Case 2: $\delta_0 \neq 0$

If $\delta_0 \neq 0$, the standardized test statistic is

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}},$$

which has approximately a $N(0,1)$ distribution if H_0 is true.

$\approx$

Comparing Two Population Proportions when $\delta_0 \neq 0$

Assumptions:

A random sample of n_i observations from population i with a proportion p_i having a characteristic of interest, and $y_i \geq 10$, $n_i - y_i \geq 10$, $i = 1, 2$.

Observed Value of Standardized Test Statistic:

$$z^* = (\hat{p}_1 - \hat{p}_2 - \delta_0) \Big/ \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$$

Approximate P-values:

H_0	H_a	p-value
$p_1 - p_2 = \delta_0$	$p_1 - p_2 < \delta_0$	$p_- = Pr(N(0,1) \leq z^*)$
$p_1 - p_2 = \delta_0$	$p_1 - p_2 > \delta_0$	$p^+ = Pr(N(0,1) \geq z^*)$
$p_1 - p_2 = \delta_0$	$p_1 - p_2 \neq \delta_0$	$p\pm = 2\min(p_-, p^+)$

$\approx$

Comparing Two Population Proportions

Example:

In a recent survey on academic dishonesty 24 of the 200 female college students surveyed and 26 of the 100 male college students surveyed agreed or strongly agreed with the statement "Under some circumstances academic dishonesty is justified." Suppose p_f denotes the proportion of all female and p_m the proportion of all male college students who agree or strongly agree with this statement.

Comparing Two Population Proportions

To illustrate the calculation of the two possible test statistics, we will consider two different scientific hypotheses:

1. **Scientific Hypothesis 1:** There is a difference in the population proportions of male and female students who agree or strongly agree with the statement.
2. **Scientific Hypothesis 2:** The population proportion of males who agree or strongly agree with the statement is at least 0.1 greater than the population proportion of females who agree or strongly agree with the statement.

Comparing Two Population Proportions

1. **Scientific Hypothesis 1** There is a difference in the population proportions of male and female students who agree or strongly agree with the statement.
2. **The Statistical Model** The two-population binomial.
3. **The Statistical Hypotheses**

$$\begin{array}{lrcl} H_0: & p_f - p_m & = & 0 \\ H_a: & p_f - p_m & \neq & 0 \end{array}$$

$\approx$

Comparing Two Population Proportions

4. **The Test Statistic** The point estimate of $p_f - p_m$ is

$$\hat{p}_f - \hat{p}_m = 24/200 - 26/100 = -0.140,$$

and the estimate of the common value of p_f and p_m under H_0 is $\hat{p} = (26 + 24)/(200 + 100) = 0.167$.
Thus,

$$z^* = \frac{24/200 - 26/100}{\sqrt{(0.1\overline{66})(0.8\overline{33})\left(\frac{1}{200} + \frac{1}{100}\right)}} = -3.067.$$

Comparing Two Population Proportions

5. **The P-Value**

   Since $y_f = 24$, $200 - y_f = 176$, $y_m = 26$, and $100 - y_m = 74$ all exceed 10, we may use the normal approximation:

   $p_{-} = P(N(0,1) \leq -3.067) = 0.0011,$

   $p^{+} = P(N(0,1) \geq -3.067) = 0.9989,$
   and

   $p\pm = 2\min(0.9989, 0.0011) = 0.0022,$

   this last being the p-value we want.

$\approx$

Comparing Two Population Proportions

1. **Scientific Hypothesis 2** The population proportion of males who agree or strongly agree with the statement is at least 0.1 greater than the population proportion of females who agree or strongly agree with the statement.
2. **The Statistical Model** The two-population binomial.
3. **The Statistical Hypotheses**

$$\begin{aligned} H_0: \quad p_f - p_m &= -0.10 \\ H_a: \quad p_f - p_m &< -0.10 \end{aligned}$$

$\approx$

Comparing Two Population Proportions

4. **The Test Statistic** The standardized test statistic is

$$\begin{aligned} z^* &= \frac{24/200 - 26/100 - (-0.10)}{\sqrt{\frac{0.12(1-0.12)}{200} + \frac{0.26(1-0.26)}{100}}} \\ &= -0.81, \end{aligned}$$

5. **The P-Value**
   $p_- = P(N(0,1) \leq -0.81) = 0.2090.$

A SAS macro to conduct these tests comparing two proportions is found here.

Fixed Significance Level Tests

We have conducted hypothesis tests by computing a p-value to measure the evidence against H_0 and in favor of H_a. When we wanted to conduct the tests at a fixed level of significance, α, we first computed the p-value and rejected H_0 if and only if the p-value was less than α.

An alternative way to conduct a test at a fixed level of significance is to determine which values of the test statistic will lead to rejection of H_0 in favor of H_a. Here are the steps involved, illustrated using the LDL reduction example:

Fixed Significance Level Tests

1. **Specify hypotheses to be tested.**

$$\begin{aligned} H_0 : \quad \mu &= 0 \\ H_a : \quad \mu &> 0 \end{aligned}$$

   (i.e. $\mu_0 = 0$)

2. **Set the significance level α.** Usual choices are 0.10, 0.01 or 0.05. We'll choose the latter.

$\approx$

Fixed Significance Level Tests

3. **Specify the (standardized) test statistic and it's distribution under $\mathbf{H}_0$.** The standardized test statistic is

$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}} = \frac{\overline{y} - 0}{15.45/\sqrt{10}},$$

and under H_0 it has a $t_{n-1} = t_9$ distribution.

$\approx$

Fixed Significance Level Tests

4. **Find the critical region of the test.** The critical region of the test is the set of values of the (standardized) test statistic for which H_0 will be rejected in favor of H_a. Here, H_a tells us that the critical region has the form

$$[t_{n-1,1-\alpha}, \infty) = [t_{9,0.95}, \infty) = [1.8331, \infty),$$

meaning H_0 will be rejected if and only if the observed value of t is greater than or equal to 1.8331.

$\approx$

Fixed Significance Level Tests

5. **Perform the test.** For the LDL example, the observed value of t is
$$t^* = \frac{21.19 - 0}{15.45/\sqrt{10}} = 4.337,$$
which falls in the critical region, so H_0 is rejected in favor of H_a.

Power of a Test

In a fixed significance level test, power is the proportion of all samples for which H_0 will be rejected in favor of H_a. Power will vary for different values of the parameter being tested, so it is written as a function of that parameter.

$\approx$

Power of a Test

Example: A random sample of size n is taken from a $N(\mu, 25)$ population. We want to test $H_0 : \mu = 10$ versus $H_a : \mu < 10$ at the 0.05 level of significance using a fixed significance level test.

The test statistic is $\overline{y}$, which under H_0 has distribution $N(10, 25/n)$, where n is the sample size.

The form of H_a tells us that small values of $\overline{y}$ should lead to rejection of H_0, which means small values of the standardized statistic $z = (\overline{y} - 10)/(5/\sqrt{n})$ should lead to rejection.

Since $Pr(z \leq z_{0.05}) = Pr(z \leq -1.645) = 0.05$, the rejection region is defined by $(\overline{y} - 10)/(5/\sqrt{n}) \leq -1.645$, which after some algebra, becomes $\overline{y} \leq 10 - (1.645)(5)/\sqrt{n}$.

$\approx$

Power of a Test

To compute the power of this test, we need to evaluate the proportion of all samples which lead to rejection when the true population mean is μ.

We'll write this as

$$\Pi(\mu) = Pr_{\mu}(\text{reject } H_0) = Pr_{\mu}(\overline{y} \leq 10 - (1.645)(5)/\sqrt{n}),$$

for all values $\mu < 10$.

$\approx$

Power of a Test

We evaluate by standardizing:

$$\begin{aligned}\Pi(\mu) &= Pr_{\mu}\left(\frac{\overline{y}-\mu}{5/\sqrt{n}} \leq \frac{10-(1.645)(5)/\sqrt{n}-\mu}{5/\sqrt{n}}\right) \\ &= Pr(z \leq \sqrt{n}(10-\mu)/5-1.645).\end{aligned}$$

For any value of μ, $\Pi(\mu)$ can be computed using online applets, statistical software, a calculator or a table of the normal distribution.

$\approx$

Power of a Test

For instance, if $n = 16$ and $\mu = 7$, we get

$$\begin{aligned}\Pi(\mu) &= Pr(z \leq \sqrt{n}(10-\mu)/5 - 1.645) \\ &= Pr(z \leq \sqrt{16}(10-7)/5 - 1.645) \\ &= Pr(z \leq 0.755) = 0.775\end{aligned}$$

Here is a plot of the power functions, $\Pi(\mu)$ for $n = 16$ and $n = 32$:

$\approx$

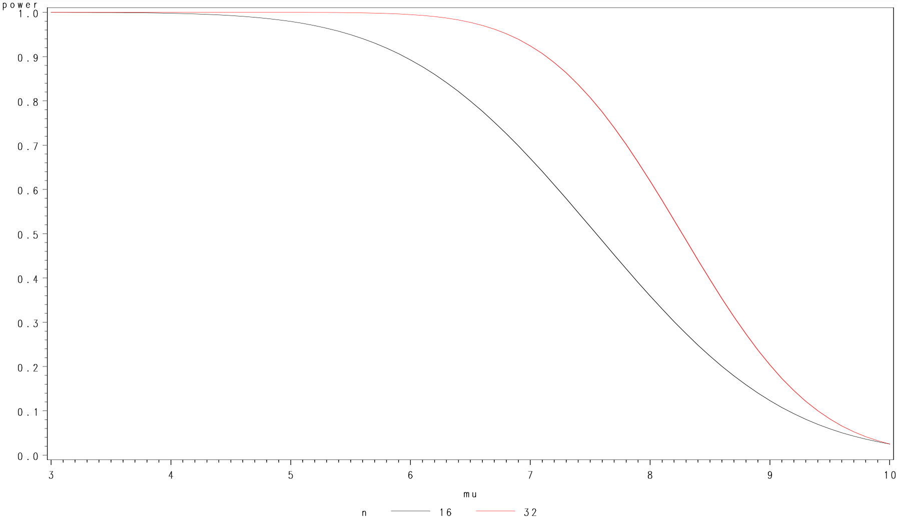

Figure: 2: Power of the one-sided test.

Power of a Test

Consider again the random sample of size n is taken from a $N(\mu, 25)$ population. We will find a formula for the rejection region of a fixed significance level test of $H_0 : \mu = 10$ versus $H_a : \mu \neq 10$ at the 0.05 level of significance. The rejection region is defined by the absolute value of the standardized test statistic exceeding a specified value.

$\approx$

Power of a Test

To define the rejection region, We need to find a value A so that $Pr(|z| \geq A) = 0.05$, where $z = (\overline{y} - 10)/(5/\sqrt{n})$.

Since $z \sim N(0, 1)$, $A = z_{0.975} = 1.96$.

Therefore, the rejection region is defined by $|(\overline{y} - 10)/(5/\sqrt{n})| \geq 1.96$, which, after some algebra, specifies rejection if and only if $\overline{y} \geq 10 + (1.96)(5)/\sqrt{n}$ or $\overline{y} \leq 10 - (1.96)(5)/\sqrt{n}$.

Power of a Test

We will now find a formula for the power function of this test. Remember the specifics:

- $\overline{y}$ is the mean of a sample of size n taken from a $N(\mu, 25)$ population.
- A level 0.05 test of $H_0 : \mu = 10$ versus $H_a : \mu \neq 10$ rejects H_0 if and only if

$$\overline{y} \geq 10 + (1.96)(5)/\sqrt{n} \text{ or } \overline{y} \leq 10 - (1.96)(5)/\sqrt{n}.$$

$\approx$

Power of a Test

The power function is

$$\begin{aligned}
\Pi(\mu) &= Pr_\mu(\text{reject } H_0) \\
&= Pr_\mu(\overline{y} \le 10 - (1.96)(5)/\sqrt{n}) + Pr_\mu(\overline{y} \ge 10 + (1.96)(5)/\sqrt{n}) \\
&= Pr(z \le \sqrt{n}(10-\mu)/5 - 1.96) \\
&+ Pr(z \ge \sqrt{n}(10-\mu)/5 + 1.96)
\end{aligned}$$

Here's a plot for $n = 16$ and 32:

$\approx$

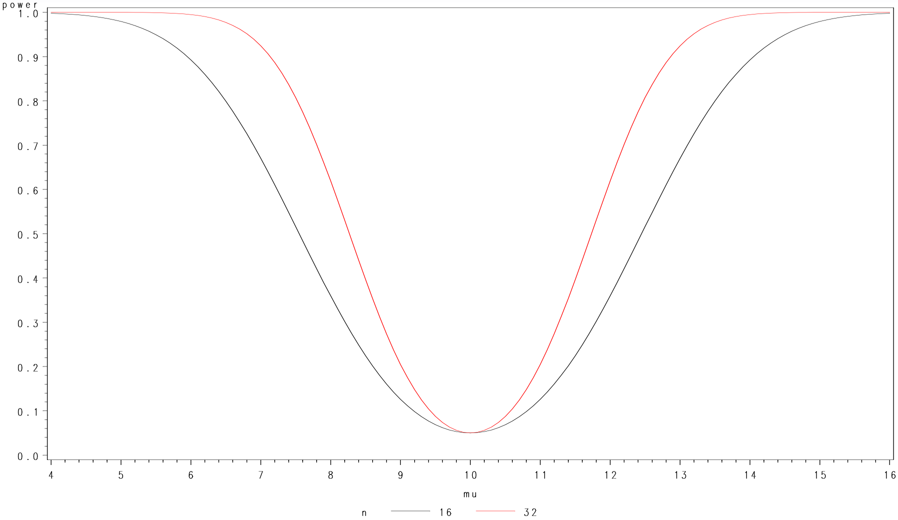

Figure: 2: Power of the two-sided test.

Power of a Test

You have seen that the power curve depends on the sample size, n. This means the power function can also be used to specify a sample size. If the researcher specifies a significance level for the test and a desired power at a specified value of the parameter being tested, then using the formulas given, (s)he can find what value of n is needed.

In our example of a test for a population mean μ, the researcher would specify a significance level α and a desired power $\Pi(\mu_0)$ at the particular mean value μ_0. A minimum sample size n to satisfy these requirements would then be obtained.

The relation between hypothesis tests and confidence intervals

There is an interesting parallel between many hypothesis tests and confidence intervals. Specifically, to each fixed-level hypothesis test we have considered, there corresponds a confidence interval that can be used to conduct the test.

$\approx$

The relation between hypothesis tests and confidence intervals

As an example, consider the following two inference procedures for a population mean μ, conducted on the same set of data:

- A two-sided level α test of $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$.
- A level $L = 1 - \alpha$ confidence interval for μ.

$\approx$

The relation between hypothesis tests and confidence intervals

The relation can be summarized in two ways:

The test will reject H_0 in favor of H_a if and only if μ_0 lies outside the confidence interval, so we can use the confidence interval to conduct the test: just construct the interval and if μ_0 is outside the interval, reject H_0 in favor of H_a.

Similarly, given the level α test of $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$, a level $L = 1 - \alpha$ confidence interval for μ consists of all μ_0 values for which the test does not reject H_0 in favor of H_a (this is called "inverting the test").

$\approx$

The relation between hypothesis tests and confidence intervals

The relation between hypothesis tests and confidence intervals enables us to do more.

First, it gives us an effective small-sample alternative to the large-sample test comparing two population proportions.

$\approx$

The relation between hypothesis tests and confidence intervals

Specifically, to test $H_0 : p_1 - p_2 = \delta_0$ versus $H_a : p_1 - p_2 \neq \delta_0$ at the α level of significance, we can construct the level $L = 1 - \alpha$ approximate score (Agresti-Coull) confidence interval

$$\tilde{p}_1 - \tilde{p}_2 \pm z_{(1+L)/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{\tilde{n}_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{\tilde{n}_2}},$$

where

$$\tilde{n}_i = n_i + 0.5z^2_{(1+L)/2},\ \tilde{p}_i = \frac{y_i + 0.25z^2_{(1+L)/2}}{\tilde{n}_i},\ i = 1, 2,$$

and reject H_0 if and only if δ_0 is not in the interval.

$\approx$

The relation between hypothesis tests and confidence intervals

Second, we can create one-sided confidence intervals by inverting tests of one-sided alternative hypotheses.

For example, consider the test of $H_0 : \mu = \mu_0$ versus $H_a : \mu > \mu_0$.

Based on what we have done, a level α test rejects H_0 in favor of H_a if

$$t_{n-1,1-\alpha} \leq \frac{\overline{y} - \mu_0}{s/\sqrt{n}} < \infty.$$

The relation between hypothesis tests and confidence intervals

After a little algebra, these inequalities are equivalent to

$$-\infty < \mu_0 \leq \bar{y} - \frac{s}{\sqrt{n}} t_{n-1,1-\alpha},$$

giving a level $L = 1 - \alpha$ confidence interval

$$\left(-\infty,\ \bar{y} - \frac{s}{\sqrt{n}} t_{n-1,1-\alpha}\right]$$

$\approx$

Recap: Hypothesis Testing

- The Components of a Statistical Hypothesis Testing Problem
    - The Scientific Hypothesis
    - The Statistical Model
    - The Statistical Hypotheses
    - The Test Statistic
    - The p-value
- Types of Hypotheses
- One and Two-Sided Tests
- The Philosophy of Hypothesis Testing
- Statistical and Practical Significance

Recap: Hypothesis Testing

- Specific Hypothesis Testing Problems:
    - 1-Sample Mean, Known Variance
    - 1 Sample Mean, Unknown Variance
    - 1-Sample Proportion, Exact
    - 1-Sample Proportion, Large Sample with Continuity Correction
    - 2-Sample Mean, Paired Observations
    - 2 Sample Mean, Pooled Variance
    - 2 Sample Mean, Separate Variance
    - 2-Sample Proportion, Large Sample
- Fixed Significance Level Tests
- Power
- Hypothesis Tests and Confidence Intervals