One-Sample Tests of Hypotheses

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy.stats import norm, probplot
```

General Concepts of Statistical Hypotheses

Often, the problem confronting the scientist or engineer is not so much the estimation of a population parameter but rather the formation of a data-based decision procedure that can produce a conclusion about some scientific system. For example,

- a medical researcher may decide on the basis of experimental evidence whether coffee drinking increases the risk of cancer in humans;
- an engineer might have to decide on the basis of sample data whether there is a difference between the accuracy of two kinds of gauges; or
- a sociologist might wish to collect appropriate data to enable him or her to decide whether a person's blood type and eye color are independent variables.

In each of these cases, the scientist or engineer *postulates* or *conjectures* something about a system. In addition, each must make use of experimental data and make a decision based on the data. In each case, the conjecture can be put in the form of a statistical hypothesis. Procedures that lead to the acceptance or rejection of statistical hypotheses such as these comprise a major area of statistical inference.

> A **statistical hypothesis** is an assertion or conjecture concerning one or more populations.

The truth or falsity of a statistical hypothesis is never known with absolute certainty unless we examine the entire population. This, of course, would be impractical in most situations. Instead, we take a random sample from the population of interest and use the data contained in this sample to provide evidence that either "*supports*" or "*does not support*" the hypothesis.

The Null and Alternative Hypotheses

The structure of hypothesis testing will be formulated with the use of the term **null hypothesis**, which refers to any hypothesis we wish to test and is denoted by H_0. The rejection of H_0 leads to the acceptance of an **alternative hypothesis**, denoted by H_1 (or sometimes as H_a). The alternative hypothesis H_1 usually represents the question to be answered or the theory to be tested, and thus its specification is crucial. The null hypothesis H_0 nullifies or opposes H_1 and is often the logical complement to H_1. As the reader gains more understanding of hypothesis testing, he or she should note that the analyst arrives at one of the two following conclusions:

> - **Reject $\boldsymbol{H_0}$** in favor of H_1 because of sufficient evidence in the data.
> - **Fail to reject $\boldsymbol{H_0}$** because of insufficient evidence in the data.

Note that the conclusions do *not* involve a formal and literal "*accept H_0.*"

The statement of H_0 often represents the "status quo" (current status) in opposition to the new idea or conjecture, stated in H_1. For example, suppose that the hypothesis postulated by the engineer is that the fraction defective p in a certain process is 0.10, the practical issue may be a concern that the historical defective probability of 0.10 no longer is true. Indeed, the conjecture may be that p exceeds 0.10. We may then state

$$\begin{aligned} H_0 &: \quad p = 0.10, \\ H_1 &: \quad p > 0.10. \end{aligned}$$

The experiment is to observe a random sample of the product in question. If 12 defective items out of 100 was observed. This does not refute $p = 0.10$, so the conclusion is "fail to reject H_0." However, if the data produce 20 out of 100 defective items, then the conclusion is "reject H_0" in favor of $H_1 : p > 0.10$. Do note that for now, the choice to decide whether or not to reject H_0 when observing a particular result is somewhat arbitrary to the engineer.

Test Procedures

A test procedure is a rule, based on sample data, for deciding whether to reject H_0. A test of $H_0 : p = 0.1$ versus $H_1 : p > 0.1$ in the defective items problem might be based on examining a random sample of items. Let X denote the number of defective items in the sample, a binomial random variable; x represents the observed value of X. If H_0 is true, $E(X) = np = 0.1(100) = 10$, whereas we can expect more than 10 defectives if H_1 is true. A value x just a bit above 10 does not strongly contradict H_0, so it is reasonable to reject H_0 only if x is substantially greater than 10. One such test procedure is to reject H_0 if $x \geq 15$ and not reject H_0 otherwise. This procedure has two constituents: (1) a test statistic, or function of the sample data used to make a decision, and (2) a rejection region consisting of those x

values for which H_0 will be rejected in favor of H_1. For the rule just suggested, the rejection region consists of $x = 15, 16, 17, 18, \ldots,$ and 100. H_0 will not be rejected if $x = 0, 1, 2, \ldots,$ or 14.

> A test procedure is specified by the following:
>
> 1. A **test statistic**, a function of the sample data on which the decision (reject H_0 or do not reject H_0) is to be based
> 2. A **rejection region** or **critical region**, the set of all test statistic values for which H_0 will be rejected
>
> The null hypothesis will then be rejected if and only if the observed or computed test statistic value falls in the rejection region.

In addition, the last number that we observe in passing into the critical region is called the **critical value**. In our illustration, the critical value is the number 14. Therefore, if $x > 14$, we reject H_0 in favor of H_1. If $x \leq 14$, we fail to reject H_0.

Errors in Hypothesis Testing

The decision procedure just described could lead to either of two wrong conclusions.

> - A **type I error** consists of rejecting the null hypothesis H_0 when it is true.
> - A **type II error** involves not rejecting H_0 when H_0 is false.

In testing any statistical hypothesis, there are four possible situations that determine whether our decision is correct or in error. These four situations are

	H_0 is true	**H_0 is false**
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

In the best of all possible worlds, test procedures for which neither type of error is possible could be developed. However, this ideal can be achieved only by basing a decision on an examination of the entire population, which we know that it is often impractical. Instead of demanding error-free procedures, we must seek procedures for which either type of error is unlikely to occur. That is, a good procedure is one for which the probability of making either type of error is small. The choice of a particular rejection region cutoff value fixes the probabilities of type I and type II errors.

> The probability of committing a type I error, also called the **level of significance**, denoted by α is given by
>
> $$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true}).$$
>
> The probability of committing a type II error, denoted by β, is given by
>
> $$\beta = P(\text{type II error}) = P(\text{Not reject } H_0 \text{ when } H_0 \text{ is false}).$$

Example A certain type of automobile is known to sustain no visible damage 25% of the time in 10-mph crash tests. A modified bumper design has been proposed in an effort to increase this percentage. Let p denote the proportion of all 10-mph crashes with this new bumper that result in no visible damage. The hypotheses to be tested are

$$\begin{aligned} &H_0 : p = 0.25 \ \ \text{(no improvement)} \\ &H_1 : p > 0.25 \ \ \text{(improvement)} \end{aligned}$$

The test will be based on an experiment involving $n = 20$ independent crashes with prototypes of the new design. Intuitively, H_0 should be rejected if a substantial number of the crashes show no damage. Consider the following test procedure:

- Test statistics: $X =$ number of crashes with no visible damage.
- Rejection region: $R = \{8, 9, 10, \ldots, 20\}$; that is, reject H_0 if $x \geq 8$.

When H_0 is true, X has a binomial distribution with parameters $n = 20$ and $p = 0.25$. Then, the probability of committing a type I error for this test is

$$\begin{aligned} \alpha &= P(X \geq 8 \text{ when } X \sim b(20, 0.25)) \\ &= 1 - B(7;\ 20, 0.25) \\ &= 1 - 0.898 = 0.102 \end{aligned}$$

That is, when H_0 is actually true, roughly 10% of all experiments consisting of 20 crashes would result in H_0 being incorrectly rejected (a type I error).

In contrast to α, there is not a single β (probability of type II error). Instead, there is a different β for each different p that exceeds .25. For example, when p = 0.3,

$$\begin{aligned} \beta(.3) &= P(\text{type II error when } p = 0.3) \\ &= P(H_0 \text{ is not rejected when } p = 0.3) \\ &= P(X \leq 7 \text{ when } X \sim b(20, 0.3)) \\ &= B(7;\ 20, 0.3) = 0.772 \end{aligned}$$

When p is actually .3 rather than $.25$ (a "small" departure from H_0), roughly 77% of all experiments of this type would result in H_0 being incorrectly not rejected!

In general, for any $0.25 < p < 1$, we have

$$\beta(p) = B(7;\ 20, p)$$

for this test procedure. Clearly, β decreases as the value of p moves farther to the right of the null value $.25$. Intuitively, the greater the departure from H_0, the less likely it is that such a departure will not be detected.

p	**.3**	**.4**	**.5**	**.6**	**.7**	**.8**
$\beta(p)$	.772	.416	.132	.021	.001	.000

Tests About a Population Mean

The general discussion in Chapter 6.2 of confidence intervals for a population mean μ focused on three different cases (variance known, large sample, and variance unknown). We now develop test procedures for these cases.

The Case of Normal Population with Variance Known

Although the assumption that the value of s is known is rarely met in practice, this case provides a good starting point. The null hypothesis in all three cases will state that μ has a particular numerical value, the null value, which we will denote by μ_0.

$$H_0 : \mu = \mu_0$$

Let $X_1, X_2, \ldots, X_n$ represent a random sample of size n from the normal population. Then the sample mean $\bar{X}$ has a normal distribution with expected value $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \sigma / \sqrt{n}$. When H_0 is true, $\mu_{\bar{X}} = \mu_0$. Consider now the statistic Z obtained by standardizing under the assumption that H_0 is true:

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

Suppose that the alternative hypothesis has the form $H_1 : \mu > \mu_0$. Then an $\bar{x}$ value less than μ_0 certainly does not provide support for H_1. Such an $\bar{x}$ corresponds to a negative value of z. Similarly, an value that exceeds μ_0 by only a small amount (corresponding to z, which is positive but small) does not suggest that H_0 should be rejected in favor of H_1. The rejection of H_0 is appropriate only when considerably exceeds μ_0—that is, when the z value is positive and large. In summary, the appropriate rejection region, based on the test statistic Z rather than $\bar{X}$, has the form $z \geq c$.

The cutoff value c should be chosen to control the probability of a type I error at the desired level of significant α. This is easily accomplished because the distribution of the test statistic Z when H_0 is true is the standard normal distribution. The required cutoff c is the z-critical value that captures upper-tail area α under the z curve. As an example, let $c = 1.645$, the value that captures tail area .05 (since $z_{0.05} = 1.645$). Then,

$$\alpha = P(Z \geq c, Z \sim N(0,1)) = P(Z > z_{0.05}) = 0.05$$

Similarly for H_1 of the form $\mu < \mu_0$ and $\mu \neq \mu_0$, we can do the same procedure but with lower-tail and two-sided tail instead. The test procedure for each case is summarized below.

> **Test Procedure for μ of Normal Population when σ is Known**
> For a test procedure with significance of α about mean μ of a normal population.

Null Hypothesis	Test Statistic	Alternative Hypothesis	Rejection Region for Level α Test
		$H_1 : \mu > \mu_0$	$z \geq z_\alpha$
$H_0 : \mu = \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$	$H_1 : \mu < \mu_0$	$z \leq -z_\alpha$
		$H_1 : \mu \neq \mu_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$

Example A manufacturer of sprinkler systems used for fire protection in office buildings claims that the true average system-activation temperature is $130\,^\circ F$. A sample of $n = 9$ systems, when tested, yields a sample average activation temperature of $131.08\,^\circ F$. If the distribution of activation times is normal with standard deviation $1.5\,^\circ F$, does the data contradict the manufacturer's claim at significance level 0.01?

Let μ be the true average system-activation temperature.

1. Null hypothesis: $H_0 : \mu = 130$ (that is $\mu_0 = 130$).
2. Alternative hypothesis: $H_1 : \mu \neq 130$ (concerning a departure from either direction of the claimed value).
3. Test statistic value:

$$z = \frac{131.08 - 130}{1.5/\sqrt{9}} = 2.16$$

4. Critical region: Since $z_{0.01/2} = z_{0.005} = 2.576$, the rejection region for two-sided test is either $z \leq -2.576$ or $z \geq 2.576$.
5. Decision: Since the computed test statistic $z = 2.16$ does not fall into the critical region ($-2.576 < z = 2.16 < 2.576$), so $\boldsymbol{H_0}$ **cannot be rejected** at significance level of 0.01. The data does not give strong support to the alternative claim.

The Case of Large-Sample

When the sample size is large, the z tests for the previous case are easily modified to yield valid test procedures without requiring either a normal population distributi. on or known σ. In the case of σ known and non-normal population, by the CLT, when the sample size n is sufficiently large (typically $n \geq 30$), we can say that

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

is approximately normal under the assumption that the population mean μ is equal to μ_0. Similarly for the case where σ is unknown, when a large-sample was used, we can approximate σ with sample standard deviation s so that

$$Z = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

is also approximately normal. Thus, we can used the same test procedure as in the case of normal population with variance known to perform a *approximately* α level test, replacing σ with s when it is required.

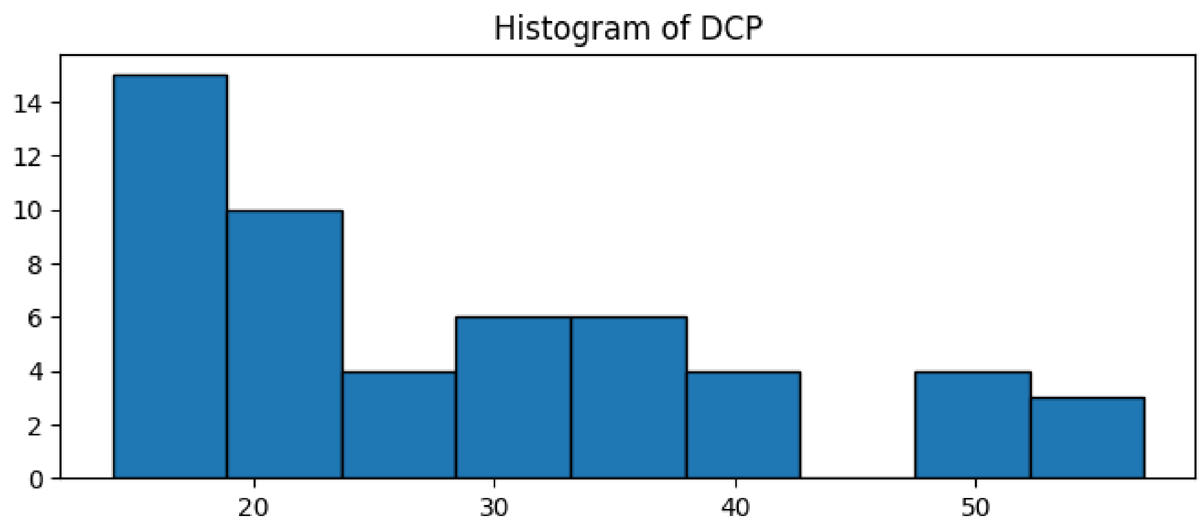
Example A dynamic cone penetrometer (DCP) is used for measuring material resistance to penetration (mm/blow) as a cone is driven into pavement or subgrade. Suppose that for a particular application it is required that the true average DCP value for a certain type of pavement be less than 30. The pavement will not be used unless there is conclusive evidence that the specification has been met. Let's state and test the appropriate hypotheses using the following data:

In [2]:
```
DCP = pd.Series([
    14.1 ,14.5 ,15.5 ,16.0 ,16.0 ,16.7 ,16.9 ,17.1 ,17.5 ,17.8,
    17.8 ,18.1 ,18.2 ,18.3 ,18.3 ,19.0 ,19.2 ,19.4 ,20.0 ,20.0,
    20.8 ,20.8 ,21.0 ,21.5 ,23.5 ,27.5 ,27.5 ,28.0 ,28.3 ,30.0,
    30.0 ,31.6 ,31.7 ,31.7 ,32.5 ,33.5 ,33.9 ,35.0 ,35.0 ,35.0,
    36.7 ,40.0 ,40.0 ,41.3 ,41.7 ,47.5 ,50.0 ,51.0 ,51.8 ,54.4,
    55.0 ,57.0
])
DCP.describe()
```

```
Out[2]: count    52.000000
        mean     28.761538
        std      12.264698
        min      14.100000
        25%      18.275000
        50%      27.500000
        75%      35.000000
        max      57.000000
        dtype: float64
```

```python
In [3]: fig, ax = plt.subplots(1,1,figsize=(8,3))

ax.hist(DCP, edgecolor="k", bins=9)
ax.set_title("Histogram of DCP");
```



The sample mean DCP is less than 30. However, there is a substantial amount of variation in the data, so the fact that the mean is less than the design specification cutoff may be a consequence just of sampling variability. Notice that the histogram does not resemble at all a normal curve, but the large-sample z tests do not require a normal population distribution.

Let μ be the true average DCP value.

1. Null hypothesis: $H_0: \mu = 30$.
2. Alternative hypothesis: $H_1: \mu < 30$ (in this setup, the pavement will not be accepted unless the null got rejected).
3. Test statistic value:

$$z = \frac{28.7615 - 30}{12.2647/\sqrt{52}} = -0.7283.$$

4. Critical region: Let's use $.05$ level of significant, so we will reject H_0 when $z \leq -z_{0.05} = -1.645$ (a lower-tailed test).
5. Decision: Since $z = -0.7283 > -1.645$, the value of the test statistic does not fall into the rejection region. Thus, we **cannot reject H_0**, so the use of the pavement is not

justify.

The Case of Normal Population in General

Another case to consider is for a normal population with σ unknown but the sample size n is not sufficiently large enough to use s to approximate σ. The distribution that can handle this situation, as discussed in Chapter 5, is the Student t-distribution. Similar to confident intervals, the test procedure can be done the same way as in previous case by using the sample deviation s to estimate σ and replace z value with t value instead.

> **Test Procedure for μ of Normal Population when σ is Unknown**
> For a test procedure with significance of α about mean μ of a normal population.

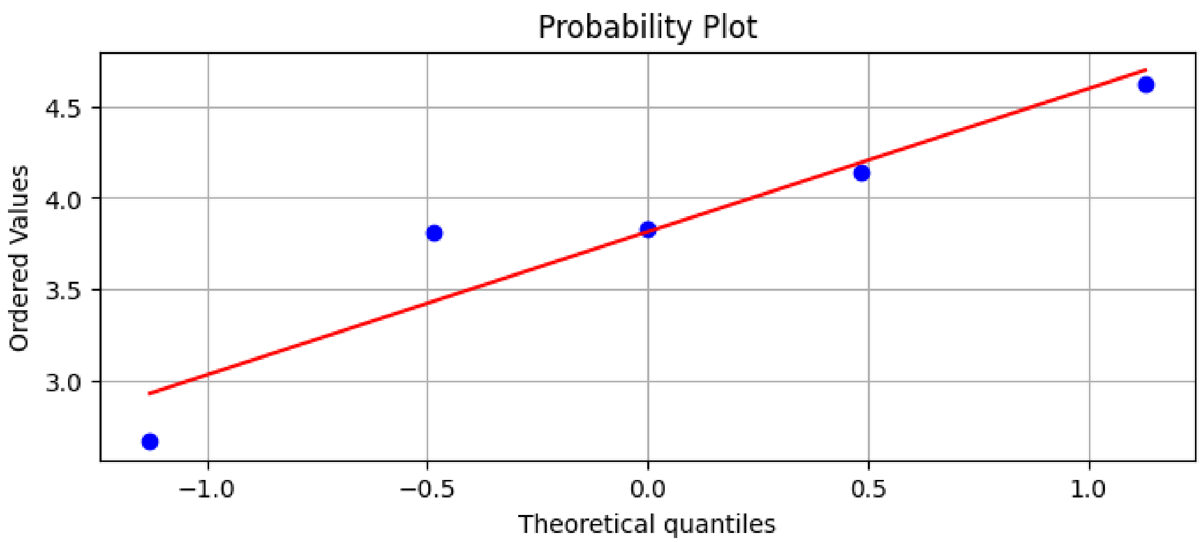
Null Hypothesis	Test Statistic	Alternative Hypothesis	Rejection Region for Level α Test
		$H_1: \mu > \mu_0$	$t \geq t_{\alpha;n-1}$
$H_0: \mu = \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$H_1: \mu < \mu_0$	$t \leq -t_{\alpha;n-1}$
		$H_1: \mu \neq \mu_0$	either $t \geq t_{\alpha/2;n-1}$ or $t \leq -t_{\alpha/2;n-1}$

Example Glycerol is a major by-product of ethanol fermentation in wine production and contributes to the sweetness, body, and fullness of wines. A particular batch of wine includes the following observations on glycerol concentration (mg/mL) for samples of standard-quality white wines: $2.67, 4.62, 4.14, 3.81, 3.83$. Suppose the desired concentration value is 4. Does the sample data suggest that true average concentration is something other than the desired value?

Even though the sample size is very small, some analysis provides strong support for assuming that the population distribution of glycerol concentration is normal. Thus, a t test can be carried out for this problem. (The probability plot Q-Q plot shown below is a tool used to determine normality in data, linear points imply normal distribution.)

In [4]:
```
fig, ax = plt.subplots(1,1,figsize=(8,3))

ax.grid()
probplot([2.67, 4.62, 4.14, 3.81, 3.83], plot=ax);
```



Let μ be the true average value of glycerol concentration (mg/mL) of standard-quality white wines.

1. Null hypothesis: $H_0 : \mu = 4$.
2. Alternative hypothesis: $H_1 : \mu \neq 4$.
3. Test statistic value: the sample mean and standard deviation is $\bar{x} = 3.814$ and $s = 0.7185$ so

$$t = \frac{3.814 - 4}{0.7185/\sqrt{5}} = -0.5788.$$

4. Critical region: Let's use .05 level of significant, so we will reject H_0 when t is outside the range $\pm t_{0.025\,;\,4}$ that is when t is outside $-2.776 < t < 2.776$ (a two-tailed test).
5. Decision: Clearly, t is outside the rejection region $(-2.776 < t = -0.5788 < 2.776)$. Therefore, it is still plausible that $\mu = 4$.

Tests About a Population Proportion

Large-Sample Tests

If $\hat{p}$ is a sample proportion and a point estimator for the population proportion p, we know that p is equivalent to the mean of a Bernoulli population. So, by the CLT,

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

has approximately a standard normal distribution if the sample size n is sufficiently large (typically $n \geq 30, np \geq 5, n(1-p) \geq 5$). Thus, given a null hypothesis about the proportion as $H_0: p = p_0$, when H_0 is true and the sample is large, we can say that

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0, 1).$$

With all these information, we can used almost the same z test procedure done for a mean.

> **Test Procedure for p with a Large-Sample**
> For a test procedure with significance of α about proportion p of a population, given that the sample size n is sufficiently large. And as a rule of thumb, the test procedure is valid when $np_0 \geq 10$ and $n(1-p_0) \geq 10$:

Null Hypothesis	Test Statistic	Alternative Hypothesis	Rejection Region for Level α Test
$H_0: p = p_0$	$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$	$H_1: p > p_0$	$z \geq z_\alpha$
		$H_1: p < p_0$	$z \leq -z_\alpha$
		$H_1: p \neq p_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$

Example Natural cork in wine bottles is subject to deterioration, and as a result wine in such bottles may experience contamination. In a tasting of commercial chardonnays, 16 of 91 bottles were considered spoiled to some extent by cork-associated characteristics. Does this data provide strong evidence for concluding that more than 15% of all such bottles are contaminated in this way? Let's carry out a test of hypotheses using a significance level of $.10$.

Let p be the true proportion of all commercial chardonnay bottles considered spoiled to some extent by cork-associated characteristics.

1. The null hypothesis is $H_0: p = 0.15$.
2. The alternative hypothesis is $H_1: p > 0.15$.
3. Test statistic value: Since $np_0 = (91)(.15) = 13.65$ and $n(1-p_0) = (91)(.85) = 77.35$ both greater than 10, the large-sample z test can be used with the test statistic value

$$z = \frac{16/91 - 0.15}{\sqrt{(0.15)(0.85)/91}} = 0.690.$$

4. For a significance of $\alpha = 0.1$, the bounds for rejecting H_0 is $z \geq z_{0.1}$, which is $z \geq 1.28$.
5. Decision: Since $0.69 < 1.28$, the test statistic does not fall into the rejection region. Thus, we cannot reject H_0 t significance level of 0.1. Although the sample proportion

$\hat{p} = 16/91 = 0.1758$ is quite larger that the assumption of $p = 0.15$ but the difference is still not enough to assert that $p > 0.15$ when taking into account the variability.

Small-Sample Tests

Test procedures when the sample size n is small are based directly on the binomial distribution rather than the normal approximation. Consider the alternative hypothesis $H_1 : p > p_0$ as an example and again let X be the number of successes in the sample. Then X is the test statistic, and the upper-tailed rejection region has the form $x \geq c$. When $H_0 : p = p_0$ is true, X has a binomial distribution with parameters n and p_0, so

$$\begin{aligned}\alpha &= P(H_0 \text{ is rejected when it is true}) \\ &= P(X \geq c \text{ when } X \sim B(n, p_0)) \\ &= 1 - B(c-1;\ n, p_0).\end{aligned}$$

So one can find a appropriate critical value c from the binomial sum table for a α level of significance test. Because X has a discrete probability distribution, it is usually not possible to find a value of c for which α is exactly the desired significance level (e.g., $.05$ or $.01$). Instead, the value of c that create the largest rejection region satisfying $P(\text{type I error}) \leq \alpha$ is used.

Or in another way, one can test a hypothesis about a proportion by computing the probability of observing a proportion *at least as extreme as* the observed sample (this probability is called P-value) value when assuming that H_0 is true, and then reject H_0 only if the probability is less than or equal to the specify significance of the test. The following example will illustrate this method.

Example A builder claims that heat pumps are installed in 70% of all homes being constructed today in the city of Richmond, Virginia. Would you agree with this claim if a random survey of new homes in this city showed that 8 out of 15 had heat pumps installed? Use a 0.10 level of significance.

Let p be the true proportion of all homes all homes being constructed today in the city of Richmond, Virginia with heat pumps installed.

1. $H_0: p = 0.70$.
2. $H_1: p \neq 0.70$.
3. Test statistic value: For $X \sim b(15, 0.70)$ with $\mu = (15)(0.70) = 10.5$, the probability of seeing x value as extreme as 8 successes is

$$2P(X \leq 8) = 2B(8; 15, 0.70) = 0.2623.$$

4. Decision: Since $0.2623 > 0.10$, we cannot reject H_0. Conclude that there is insufficient reason to doubt the builder's claim.

Tests About a Population Variance

Using the fact that for a sample of size n from a normal population with variance σ^2,

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

has a chi-squared distribution with $v = n - 1$ degrees of freedom. So we can used the statistic χ^2 for test procedures concerning a variance. Consider a null hypothesis of the form $H_0: \sigma^2 = \sigma_0^2$ such that when H_0 is true,

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(v = n - 1)$$

We can derived the test procedure for α level tests as summarized below.

> **Test Procedure for σ^2 of a Normal Population**
> For a test procedure with significance of α about variance σ^2 of a normal population:

Null Hypothesis	Test Statistic	Alternative Hypothesis	Rejection Region for Level α Test
$H_0: \sigma^2 = \sigma_0^2$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$H_1: \sigma^2 > \sigma_0^2$	$\chi^2 \geq \chi^2_{\alpha:(n-1)}$
		$H_1: \sigma^2 < \sigma_0^2$	$\chi^2 \leq \chi^2_{1-\alpha;(n-1)}$
		$H_1: \sigma^2 \neq \sigma_0^2$	either $\chi^2 \geq \chi^2_{\alpha/2;(n-1)}$ or $\chi^2 \leq \chi^2_{1-\alpha/2;(n-1)}$

Example A manufacturer of car batteries claims that the life of the company's batteries is approximately normally distributed with a standard deviation equal to 0.9 year. If a random sample of 10 of these batteries has a standard deviation of 1.2 years, do you think that $\sigma > 0.9$ year? Use a 0.05 level of significance.

Let σ be the true standard deviation of batteries life of the manufacturer we are interested in.

1. Null hypothesis: Since the assumption is $\sigma = 0.9$. Thus, $H_0 : \sigma^2 = 0.81$.
2. Alternative hypothesis: $H_1 : \sigma^2 > 0.81$.
3. Test statistic value: From the given sample standard deviation $s = 1.2$, the standard variance is then $s^2 = 1.44$. So

$$\chi^2 = \frac{(9)(1.44)}{0.81} = 16.$$

4. Rejection region: For a significance level of 0.05, the value of $\chi^2_{0.05\,;\,(9)}$ is 16.919. Thus, we reject H_0 when $\chi^2 > 16.919$.
5. Decision: We can see that $\chi^2 = 16 < 16.919$, so we cannot reject H_0 at the significance level of 0.05. So the statistic χ^2 does not give a strong evident that $\sigma > 9$, however, the statistic is being very close to begin within the rejection region. In fact, some further test (such as P-value) may arise an assertion that $\sigma > 0.9$, so the analyst should approach the use of this particular χ^2-test with caution.

Two-Samples Tests of Hypotheses

Tests About Difference Between Two Means

The general discussion in Chapter 6.2 of confidence intervals for a population mean μ focused on three different cases (variance known, large sample, and variance unknown). We now develop test procedures for these cases.

The reader should now understand the relationship between tests and confidence intervals, and can only heavily rely on details supplied by the confidence interval material in Chapter 6.2.

Two independent random samples of sizes n_1 and n_2, respectively, are drawn from two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. We know that the random variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

has a standard normal distribution. Here we are assuming that n_1 and n_2 are sufficiently large that the Central Limit Theorem applies or when the two populations are normal for small n_1 and n_2.

Now, consider a test based on a hypothesis

$$H_0 : \mu_1 - \mu_2 = d_0$$

Suppose the values $\bar{x}_1$ and $\bar{x}_2$ are computed from samples and, for σ_1^2 and σ_2^2 known, the test statistic is given by

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

with a two-tailed critical region in the case of a two-sided alternative. That is, reject H_0 in favor of $H_1 :\ \mu_1 - \mu_2 = d_0$ if $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$. One-tailed critical regions are used in the case of the one-sided alternatives. The reader should, as before, study the test statistic and be satisfied that for, say, $H_1 :\ \mu_1 - \mu_2 > d_0$, the signal favoring H_1 comes from large values of z. Thus, the upper-tailed critical region $z > z_\alpha$ applies.

> **Test Procedure for $\mu_1 - \mu_2$ of Normal Populations when σ_1, σ_2 are Known**
> For a test procedure with significance of α about difference between two means $\mu_1 - \mu_2$ of two independent normal populations with known variances σ_1^2 and σ_2^2 respectively.

Null Hypothesis	Test Statistic	Alternative Hypothesis	Rejection Region for Level α Test
$H_0 : \mu_1 - \mu_2 = d_0$	$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$	$H_1 : \mu_1 - \mu_2 > d_0$	$z \geq z_\alpha$
		$H_1 : \mu_1 - \mu_2 < d_0$	$z \leq -z_\alpha$
		$H_1 : \mu_1 - \mu_2 \neq d_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$

The Case of Unknown but Equal Variances

The more prevalent situations involving tests on two means are those in which variances are unknown. If the scientist involved is willing to assume that both distributions are normal and that $\sigma_1 = \sigma_2 = \sigma$, the pooled t-test (often called the two-sample t-test) may be used. The test statistic is given by the following test procedure.

> **Test Procedure for $\mu_1 - \mu_2$ of Normal Populations when σ_1, σ_2 are Unknown but Equal (Two Samples Pooled t-test)**
>
> For a test procedure with significance of α about difference between two means $\mu_1 - \mu_2$ of two independent normal populations with unknown but equal variances. We'll be using a t-test with $v = n_1 + n_2 - 2$ degrees of freedom. Given a pooled estimation for the variance
>
> $$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$
>
> | Null Hypothesis | Test Statistic | Alternative Hypothesis | Rejection Region for Level α Test |
> |---|---|---|---|
> | $H_0 : \mu_1 - \mu_2 = d_0$ | $t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$ | $H_1 : \mu_1 - \mu_2 > d_0$ | $t \geq t_\alpha$ |
> | | | $H_1 : \mu_1 - \mu_2 < d_0$ | $t \leq -t_\alpha$ |
> | | | $H_1 : \mu_1 - \mu_2 \neq d_0$ | either $t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$ |

Example An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 85 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 81 with a sample standard deviation of 5. Can we conclude at the 0.05 level of significance that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units? Assume the populations to be approximately normal with equal variances.

Let μ_1 and μ_2 be the true population means of the abrasive wear for material 1 and material 2 respectively.

1. Setting hypotheses:

$$H_0 : \mu_1 - \mu_2 = 2,$$
$$H_1 : \mu_1 - \mu_2 > 2.$$

2. Choose level of significance: $\alpha = 0.05$
3. Critical region: Since we will use t-test with $v = 12 + 10 - 2 = 20$ degrees of freedom. So we will reject H_0 in favor of H_1 when $t > t_{0.05;(20)} = 1.725$.
4. Computation:

$$s_p^2 = \frac{(11)(4^2) + (9)(5^2)}{12 + 11 - 2} = 20.05$$

$$t = \frac{(85 - 81) - 2}{\sqrt{(20.05)(1/12 + 1/10)}} = 1.04$$

5. Decision: Do not reject H_0 since $t = 1.04 < 1.725$. 0. We are unable to conclude that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units.

The Case of Unknown and Unequal Variances

There are situations where the analyst is not able to assume that $\sigma_1 = \sigma_2. If the populations are normal, the statistic

$$T = \frac{(\mu_1 - \mu_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

has an approximate t-distribution with approximate degrees of freedom

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

As a result, we obtain the test procedure:

> **Test Procedure for $\mu_1 - \mu_2$ of Normal Populations when σ_1, σ_2 are Unknown and Unequal**
> For a test procedure with significance of α about difference between two means $\mu_1 - \mu_2$ of two independent normal populations with unknown variances. We'll be using a t-test with degrees of freedom given by
>
> $$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$$
>
> | Null Hypothesis | Test Statistic | Alternative Hypothesis | Rejection Region for Level α Test |
> | --- | --- | --- | --- |
> | | | $H_1: \mu_1 - \mu_2 > d_0$ | $t \geq t_\alpha$ |
> | $H_0: \mu_1 - \mu_2 = d_0$ | $t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$ | $H_1: \mu_1 - \mu_2 < d_0$ | $t \leq -t_\alpha$ |
> | | | $H_1: \mu_1 - \mu_2 \neq d_0$ | either $t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$ |

Paired Observations

Testing of two means can be accomplished when data are in the form of paired observations. In this pairing structure, the conditions of the two populations (treatments) are assigned randomly within homogeneous units. Computation of the confidence interval for $\mu_1 - \mu_2$ in the situation with paired observations is based on the random variable

$$T = \frac{\bar{D} - \mu_D}{S_d/\sqrt{n}},$$

where $\bar{D}$ and S_d are random variables representing the sample mean and standard deviation of the differences of the observations in the experimental units. As in the case of the pooled t-test, the assumption is that the observations from each population are normal. This two-sample problem is essentially reduced to a one-sample problem by using the computed differences $d_1, d_2, \ldots, d_n$. Thus, the hypothesis reduces to $H_0 : \mu_D = d_0$.

The computed test statistic is then given by

$$t = \frac{\bar{d} - d_0}{s_d/\sqrt{n}}.$$

> **Test Procedure for $\mu_1 - \mu_2$ of Paired Observations**
> If $\bar{d}$ and s_d are the mean and standard deviation, respectively, of the normally distributed differences of n random **pairs of measurements**. Then the test procedure for the real mean μ_D $(= \mu_1 - \mu_2)$ of population of differences is a t-test with $n - 1$ degrees of freedom given by:
>
> | Null Hypothesis | Test Statistic | Alternative Hypothesis | Rejection Region for Level α Test |
> |---|---|---|---|
> | | | $H_1 : \mu_d > d_0$ | $t \geq t_\alpha$ |
> | $H_0 : \mu_d = d_0$ | $t = \frac{\bar{d} - d_0}{s_d/\sqrt{n}}$ | $H_1 : \mu_d < d_0$ | $t \leq -t_\alpha$ |
> | | | $H_1 : \mu_d \neq d_0$ | either $t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$ |

Example **Blood Sample Data**: In a study conducted in the Forestry and Wildlife Department at Virginia Tech, J. A. Wesson examined the influence of the drug succinylcholine on the circulation levels of androgens in the blood. Blood samples were taken from wild, free-ranging deer immediately after they had received an intramuscular injection of succinylcholine administered using darts and a capture gun. A second blood sample was obtained from each deer 30 minutes after the first sample, after which the deer was released.

The levels of androgens at time of capture and 30 minutes later, measured in nanograms per milliliter (ng/mL), for 15 deer are given in the table shown below.

No description has been provided for this image

Assuming that the populations of androgen levels at time of injection and 30 minutes later are normally distributed, test at the 0.05 level of significance whether the androgen concentrations are altered after 30 minutes.

Let μ_1 and μ_2 be the average androgen concentration at the time of injection and 30 minutes later, respectively, and μ_D be the average of difference $\mu_1 - \mu_2$.

1. $H_0 : \mu_1 - \mu_2 = 0 \iff \mu_d = 0.$
2. $H_1 : \mu_1 - \mu_2 \neq 0 \iff \mu_d \neq 0.$
3. $\alpha = 0.05.$
4. Critical region: Since we'll be using a two-tailed t-test with $v = n - 1 = 14$ degrees of freedom. Reject H_0 when $t < -t_{0.025\,;\,(14)} = -2.145$ or $t > t_{0.025\,;\,(14)} = 2.145.$
5. Computation:

In [5]:
```python
androgen_diff = np.array([
    4.26, -2.08, 2.76, 0.94, 1.11,
    3.21, 7.31, 13.74, 0.52, -2.45,
    -0.68, -0.16, 68.03, 26.55, 24.66
])

print("Sample differences mean (d bar) =", androgen_diff.mean())
print("Sample differences std (s_d) =", androgen_diff.std(ddof=1))
```

```
Sample differences mean (d bar) = 9.848
Sample differences std (s_d) = 18.473626838588807
```

Therefore

$$t = \frac{9.848 - 0}{18.474/\sqrt{15}} = 2.06.$$

6. Decision: Since $-2.145 < t = 2.06 < 2.145$, we do not reject H_0 and hence the data shows some evidence that there is a difference in mean circulating levels of androgen.

Tests About Difference Between Two Proportions

In our construction of confidence intervals for p_1 and p_2 we noted, for n_1 and n_2 sufficiently large, that the point estimator $\hat{P}_1 - \hat{P}_2$ was approximately normally distributed with mean

$$\mu_{\hat{P}_1 - \hat{P}_2} = p_1 - p_2$$

and variance

$$\sigma^2_{\hat{P}_1-\hat{P}_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

Therefore, our critical region(s) can be established by using the standard normal variable

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

To compute a value of Z, however, we must estimate the parameters p_1 and p_2 that appear in the radical. Suppose that the sample sizes are large enough to estimate the p_1 and p_2 with $\hat{P}_1$ and $\hat{P}_2$ respectively and based on a hypothesis $H_0 : p_1 - p_2 = d_0$, the computed test statistic is then

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - d_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}.$$

In a special case where $d_0 = 0$, that is when the null hypothesis said $p_1 = p_2 = p \iff p_1 - p_2 = 0$, the statistic Z become

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

where we can used pooled estimation to estimate the value of the parameter p using

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

where x_1 and x_2 are the numbers of successes in each of the two samples. Thus the computed statistic become

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$$

So we can write a test procedure for the difference of two proportion as follows.

> **Test Procedure for $p_1 - p_2$ (General Test, Large Samples)**
> For a test procedure with significance of α about difference between two proportions $p_1 - p_2$ of two independent normal populations, or non-normal

populations but with large sample sizes n_1 and n_2.

Null Hypothesis	Test Statistic	Alternative Hypothesis	Rejection Region for Level α Test
$H_0: p_1 - p_2 = d_0$	$z = \dfrac{(\hat{p}_1 - \hat{p}_2) - d_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$	$H_1: p_1 - p_2 > d_0$	$z \geq z_\alpha$
		$H_1: p_1 - p_2 < d_0$	$z \leq -z_\alpha$
		$H_1: p_1 - p_2 \neq d_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$

Test Procedure for $p_1 - p_2$ (Test for Equal Proportions, Large Samples)

For a test procedure with significance of α about difference between two proportions $p_1 - p_2$ when testing whether two proportions are equal. Given a pooled estimate for $p_1 = p_2 = p$ as

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

Null Hypothesis	Test Statistic	Alternative Hypothesis	Rejection Region for Level α Test
$H_0: p_1 - p_2 = 0$	$z = \dfrac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$	$H_1: p_1 - p_2 > 0$	$z \geq z_\alpha$
		$H_1: p_1 - p_2 < 0$	$z \leq -z_\alpha$
		$H_1: p_1 - p_2 \neq 0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$

Example A vote is to be taken among the residents of a town and the surrounding county to determine whether a proposed chemical plant should be constructed. The construction site is within the town limits, and for this reason many voters in the county believe that the proposal will pass because of the large proportion of town voters who favor the construction. To determine if there is a significant difference in the proportions of town voters and county voters favoring the proposal, a poll is taken. If 120 of 200 town voters favor the proposal and 240 of 500 county residents favor it, would you agree that the proportion of town voters favoring the proposal is higher than the proportion of county voters? Use an $\alpha = 0.05$ level of significance.

Let p_1 and p_2 be the true proportions of voters in the town and county, respectively, favoring the proposal.

1. $H_0: p_1 - p_2 = 0$.

2. $H_1 : p_1 - p_2 > 0$.
3. $\alpha = 0.05$.
4. Critical Region: $z > z_{0.05} = 1.645$.
5. Computation:

$$\hat{p}_1 = \frac{120}{200} = 0.600, \hat{p}_2 = \frac{240}{500} = 0.480 \quad \text{and} \quad \hat{p} = \frac{120 + 240}{200 + 500} = 0.514.$$

Therefore,

$$z = \frac{0.600 - 0.480}{\sqrt{(0.514)(0.486)(1/200 + 1/500)}} = 2.870$$

6. Decision: Reject H_0 in favor of H_1 since $z = 2.870 > 1.645$. Thus, the given data agree that the proportion of town voters favoring the proposal is higher than the proportion of county voters.

Tests About Ratio Between Two Variances

Now let us consider the problem of testing the **equality** of the variances σ_1^2 and σ_2^2 of two populations. That is, we shall test the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ against one of the usual alternatives

$$\sigma_1^2 < \sigma_2^2, \sigma_1^2 > \sigma_2^2, \quad \text{or} \quad \sigma_1^2 \neq \sigma_2^2.$$

For independent random samples of sizes n_1 and n_2, respectively, from the two populations, the f-value for testing $\sigma_1^2 = \sigma_2^2$ is the ratio

$$f = \frac{s_1^2}{s_2^2},$$

where s_1^2 and s_2^2 are the variances computed from the two samples. If the two populations are approximately normally distributed and the null hypothesis is true, the ratio f is a value of the F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom. Therefore, the critical regions of significance α is given as follows.

> **Test Procedure for σ_1^2/σ_2^2 of Normal Populations**
> For a test procedure with significance of α about whether two variances σ_1^2 and σ_2^2 are equal by determining whether $\sigma_1^2/\sigma_2^2 = 1$. We can used f-test with $v_1 = n_1 - 1$ and $v_2 = n_2 - 2$ degrees of freedom.

Null Hypothesis	Test Statistic	Alternative Hypothesis	Rejection Region for Level α Test
		$H_1: \sigma_1^2 > \sigma_2^2$	$f \geq f_\alpha(v_1, v_2)$
$H_0: \sigma_1^2 = \sigma_2^2$	$f = \frac{s_1^2}{s_2^2}$	$H_1: \sigma_1^2 < \sigma_2^2$	$f \leq f_{1-\alpha}(v_1, v_2)$
		$H_1: \sigma_1^2 \neq \sigma_2^2$	either $f \geq f_{\alpha/2}(v_1, v_2)$ or $f \leq f_{1-\alpha/2}(v_1, v_2)$

Example In testing for the difference in the abrasive wear of the two materials in previously mentioned example. Recall that the standard deviation for sample of 12 pieces of material 1 is 4 and for 10 pieces of material 2 is 5. We assumed that the two unknown population variances were equal. Were we justified in making this assumption? Use a 0.10 level of significance.

Let $\sigma_1^2 and$\sigma^2_2$ be the population variances for the abrasive wear of material 1 and material 2, respectively.

1. $H_0: \sigma_1^2 = \sigma_2^2$.
2. $H_1: \sigma_1^2 \neq \sigma_2^2$.
3. $\alpha = 0.10$.
4. Critical region: Reject H_0 when $f > f_{0.05}(11, 9) = 3.100$ or when $f < f_{0.95}(11, 9) = 1/f_{0.05}(9, 11) = 1/2.90 = 0.345$.
5. Computation: $f = 4^2/5^2 = 16/25 = 0.640$.
6. Decision: Since $0.345 < f = 0.640 < 3.100$, we cannot reject H_0. Conclude that there is insufficient evidence that the variances differ.