START WED 11/3/21

HANDOUT #12: HYPOTHESES TESTING

I. Principles of Testing

1. Selection of Null and Alternative Hypotheses
2. Type I and Type II Errors
3. Power of Test
4. Level of Significance
5. Significance Probability (p-value)
6. Sample Size Determination
7. Effect Size
8. Practical vs Statistical Significance

II. Tests About a Single Population/Process Parameter

1. Tests about Population Mean
   a. Normal distribution with known σ
   b. Normal distribution with σ unknown
   c. Non-normal distributions with large sample size n
   d. Exponential Distribution
   e. Poisson Distribution
2. Tests about Population Median
3. Tests about Population Proportion
4. Tests about Population Standard Deviation
5. Tests about Autocorrelation ρ_k
6. Inverting a Test to Obtain C.I.
7. Bootstrap Tests of hypotheses

Supplemental Reading:

- Section 6.3, Chapter 7, Sections 9.1 and 14.1 in Tanhane/Dunlop book

General Steps in Hypotheses Testing

Long distance runners have contended that moderate exposure to ozone increases lung capacity. To investigate the potential of this theory, a researcher selected 40 adult rats of the same age, health, and strain. She then exposed the rats to ozone at a rate of 2 ppm for a period of 30 days. The lung capacity of the rats was determined at the beginning of the study and again after 30 days of ozone exposure. She wanted to determine if the mean lung capacity increased over the 30 day period. The following two hypotheses were formulated:

Null Hypothesis: The mean lung capacity was not increased.

Alternative Hypothesis: The mean lung capacity was increased.

If the mean change in lung capacity was a large positive value, then we would reject the null hypothesis and conclude that the evidence in the data supported the alternative hypothesis.

A formal structure for testing hypotheses about a population/process parameter θ is as follows.

1. Let Θ be the parameter space of θ which we partition into Θ_o and Θ_1, such that $\Theta = \Theta_o \cup \Theta_1$ and $\Theta_o \cap \Theta_1 = \emptyset$.

2. We want to test
$$H_o : \theta \epsilon \Theta_o \quad \text{versus} \quad H_1 : \theta \epsilon \Theta_1.$$
We designate H_o the **Null hypothesis** and H_1 the **Alternative hypothesis** (often referred to as the **Research hypothesis**).

3. We collect data from the population/process $X_1, \cdots, X_n$ and construct a **test statistic,** $T(X)$ which is a function of the data and θ to evaluate which of the two hypotheses is most believable based on the value of the test statistic.
The test statistic is similar to the pivot we used in constructing confidence intervals.

4. We then find a region, **R**, of the possible values of $T(X)$, called the **Rejection Region**.
If $T(X) \in R$ we reject the null hypothesis, H_o,
If $T(X) \notin R$, we do not reject the null hypothesis, H_o:
$$\begin{aligned} T(X) \in R &\Rightarrow \text{reject } H_o \\ T(X) \notin R &\Rightarrow \text{fail to reject } H_o \end{aligned}$$

In many situations, the rejection region can be formulated as
$$R = \{X : T(X) > C\}$$
where C is called the critical value of the test. The theoretical problem is then to find the distribution of the test statistic $T(X)$ and the value of C to satisfy specified operational conditions on the testing procedure.

5. Because hypotheses testing is based on a limited amount of information from a population/process, there is always a chance that the decision reached based on the observed data is incorrect.

   We structure the assignment of values in the parameter space such that we retain H_o unless there is strong evidence in the data to reject H_o. This idea is often referred to as the Neyman-Pearson philosophy of hypotheses testing.

6. There are two types of errors that can occur.

   A **Type I error** occurs when the decision based on the data is to reject H_o when in fact H_o is true.

   A **Type II error** occurs when the decision based on the data is to fail to reject H_o when in fact H_o is false.

   We summarize the types of decisions reached in hypotheses testing in the following table:

	Decision Based on Data	
Population	Fail to Reject H_o	Reject H_o
H_o True	Correct Decision	Type I error
H_1 True	Type II error	Correct Decision

   The probability that the test makes a Correct Decision or a Type I error or a Type II error are obtained from the **power function**:

7. The **power function** of a test of hypotheses

$$H_o : \theta \epsilon \Theta_o \quad \text{versus} \quad H_1 : \theta \epsilon \Theta_1.$$

   with rejection region R is given by

$$\gamma(\theta^*) = P_{\theta^*}(T(X) \in R) = P_{\theta^*}(\text{Test Rejects } H_o) = \text{probability Test rejects } H_o \text{ when } \theta = \theta^*$$

   For any $\theta^* \in \Theta_o$, $\gamma(\theta^*)$ is the probability of a Type I error.

   For any $\theta^* \in \Theta_1$, $\gamma(\theta^*)$ is the probability of a correct decision.

   For any $\theta^* \in \Theta_1$, $1 - \gamma(\theta^*)$ is the probability of a Type II error.

   The **size** of a test of hypotheses is defined to be

$$\alpha = \sup_{\theta \in \Theta_o} \gamma(\theta)$$

   Thus, α, the size of the test, is the maximum probability of making a Type I error.

Decisions Based on a Test of Hypotheses

When conducting a test of hypotheses, the decision reached may be correct, a Type I error, or a Type II error. The consequences of these decisions are of differing severity.

For example, suppose a patient comes into a doctor's office with a set of symptoms which may indicate a disease. The doctor orders a medical test and based on the result of the test must decide between the following two hypotheses:

$$H_o : \text{Patient does not have disease} \quad \text{versus} \quad H_1 : \text{Patient has disease}$$

- A **Type I Error** (False Positive) would result if H_o is true but test indicates H_1 is true, that is,

  a patient without the disease had a test result indicating that he/she had the disease. Thus the patient unnecessarily would be given the treatment for the disease. This may cause the patient discomfort, or severe problems if drug has side effects, and financial loss.

- A **Type II Error** (False Negative) would result if H_1 is true but test indicates H_o is true, that is,

  a patient with the disease had a test result indicating that he/she did not have the disease. Thus, the patient would not receive the treatment and may have severe physical consequences.

- The goal is to carry out tests of hypotheses in a manner which has the probability of both Type I and Type II errors kept at a minimum. Unfortunately, for a finite sample size, the probabilities of Type I and Type II errors cannot be controlled simultaneously. Adjusting the constant C in the rejection region to reduce one of the probabilities results in an increase in the other probability.

- It is thus customary to assign an upper bound on the probability of Type I error and to then attempt to minimize the probability of Type II error. This strategy is referred to as the Neyman-Pearson Philosophy of hypotheses testing.

  (Is a Type I error always more consequential than a Type II error?)

- The upper bound on the probability of Type I error is referred to as the **size** or **level of significance** of the test.

  We thus have the following problem to solve:

1. Select a test statistic $\mathbf{T}(\mathbf{X})$ and a rejection region $\mathbf{R}$ such that

$$\sup_{\theta \in \Theta_o} \gamma(\theta) = \sup_{\theta \in \Theta_o} P_\theta[\text{reject} \quad H_0] = \sup_{\theta \in \Theta_o} P_\theta[\text{Type I error}] = \alpha$$

2. Subject to the above constraint, select the test statistic $T(X)$ and R to maximize $\gamma(\theta)$ for $\theta \in \Theta_1$, where

$$\gamma(\theta) = P_\theta[\text{reject} \quad H_0] = P_\theta[\text{correct decision}]$$

3. That is, to minimize $1 - \gamma(\theta)$ for $\theta \in \Theta_1$, where

$$1 - \gamma(\theta) = P_\theta[\text{fail to reject} \quad H_0] = P_\theta[\text{Type II error}]$$

In your theory courses, STAT 611 or STAT 630, theorems will be given that yield tests having maximum power for all values of the parameter falling in H_1 in comparison to all tests having size α.

These tests are called uniformly most powerful (UMP) tests. A UMP test has the largest possible power for all $\theta \epsilon \Theta_1$ when compared to any other test of the same size.

In many situations, these optimal tests are difficult to find or may not even exist. A discussion of such tests will be left for your theory courses. We will consider some general methods for determining tests of hypotheses which in some cases may yield uniformly most powerful tests.

Hypotheses are generally of two forms:

1. $\theta = \theta_o$ is called a **simple hypothesis** and

2. $\theta < \theta_o, \quad \theta > \theta_o, \quad \theta \leq \theta_o, \quad \theta \geq \theta_o, \quad \theta \neq \theta_o$ are called **composite hypotheses**.

3. Hypotheses of the form
$$H_o : \theta = \theta_o \quad \text{versus} \quad H_1 : \theta \neq \theta_o$$
are called **two-sided** hypotheses.

4. Hypotheses of the form
$$H_o : \theta \leq \theta_o \quad \text{versus} \quad H_1 : \theta > \theta_o$$
or
$$H_o : \theta \geq \theta_o \quad \text{versus} \quad H_1 : \theta < \theta_o$$
are called **one-sided** hypotheses.

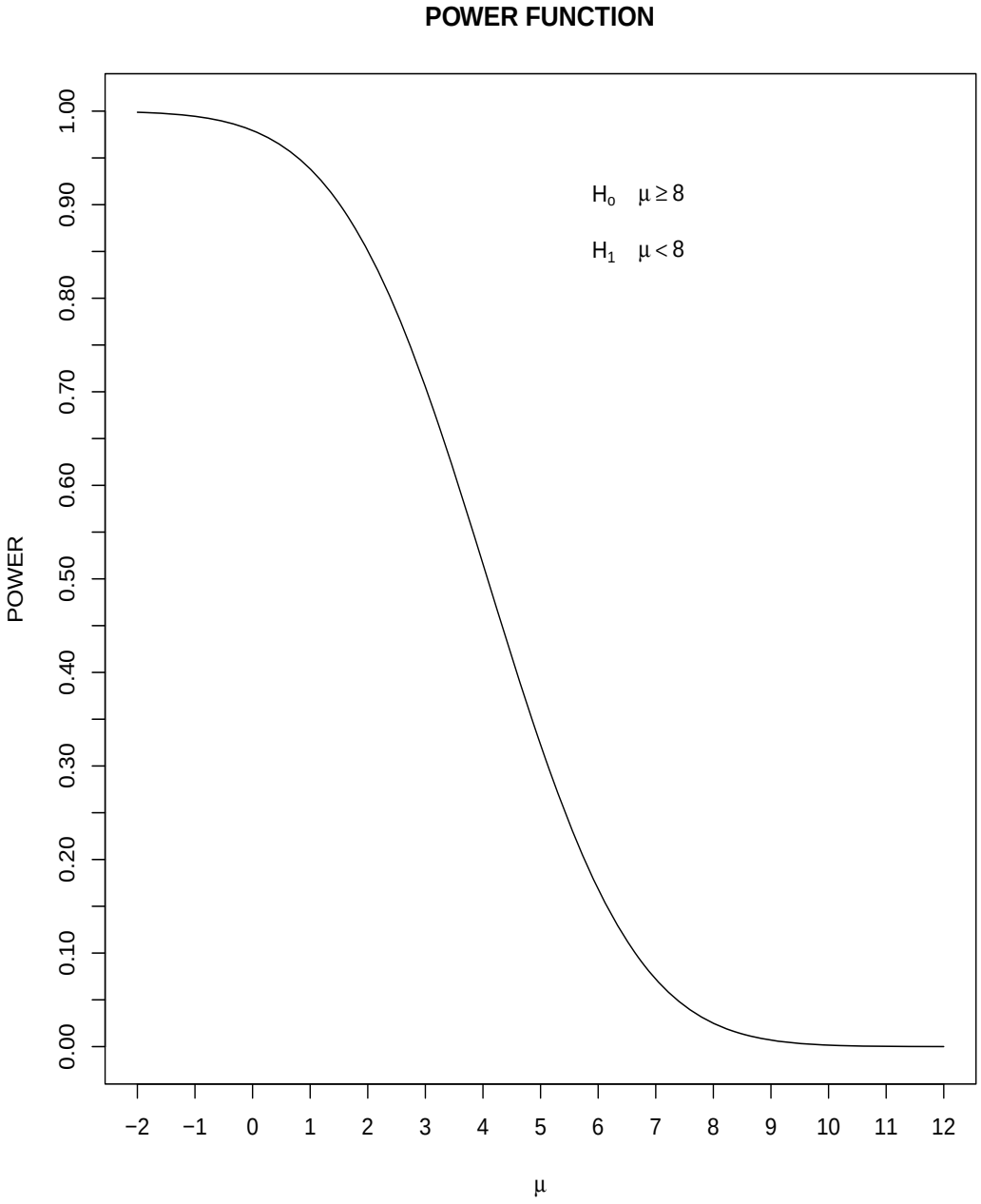
In all such hypotheses, we would determine the appropriate Rejection Region: R, The values of the test statistic $T(X)$ which imply reject H_o

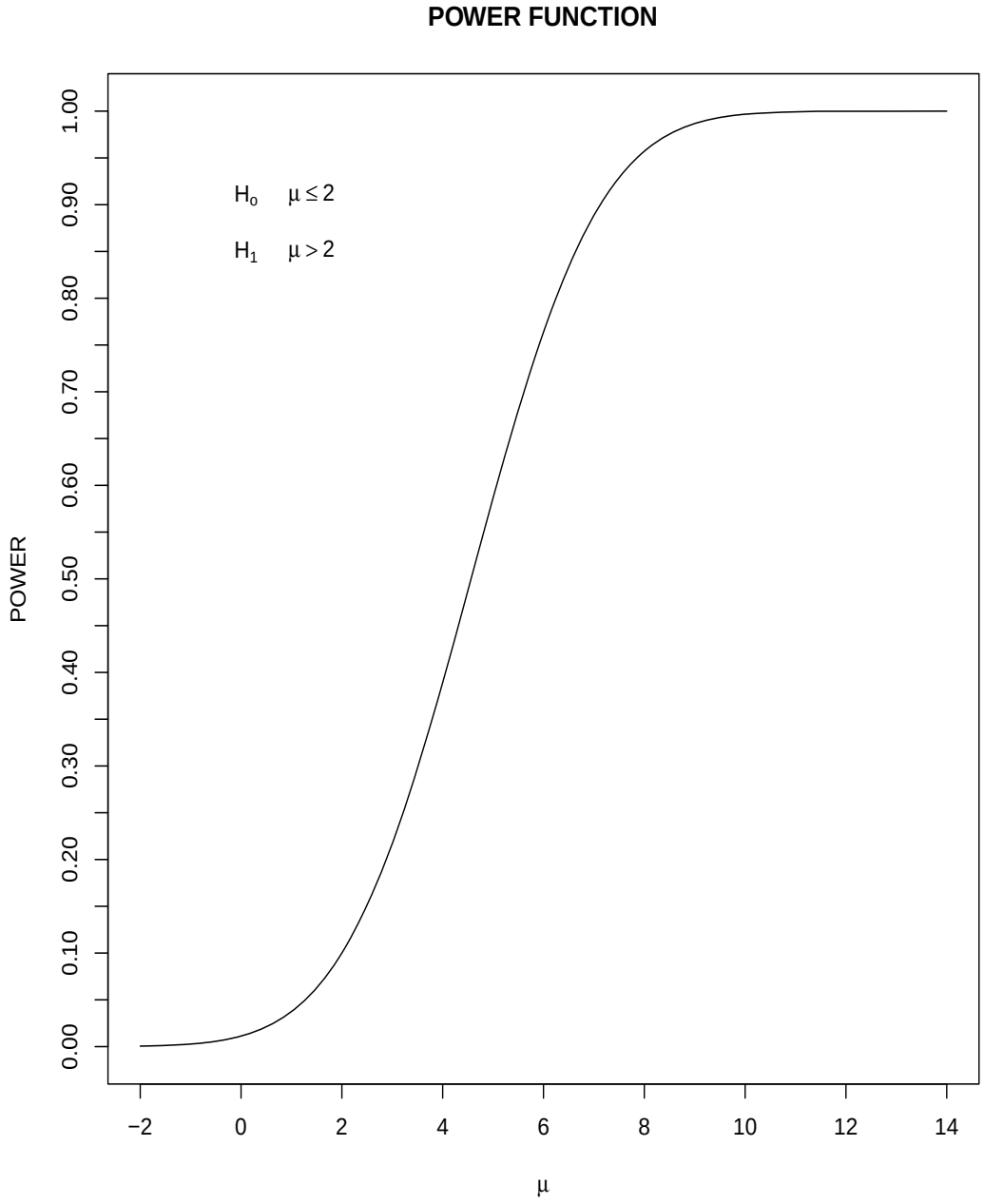
H1. For the one-sided hypotheses :
$$H_o : \theta \geq \theta_o \ \text{ versus } \ H_1 : \theta < \theta_o \qquad R = \{X : T(X) < C_\alpha\}$$

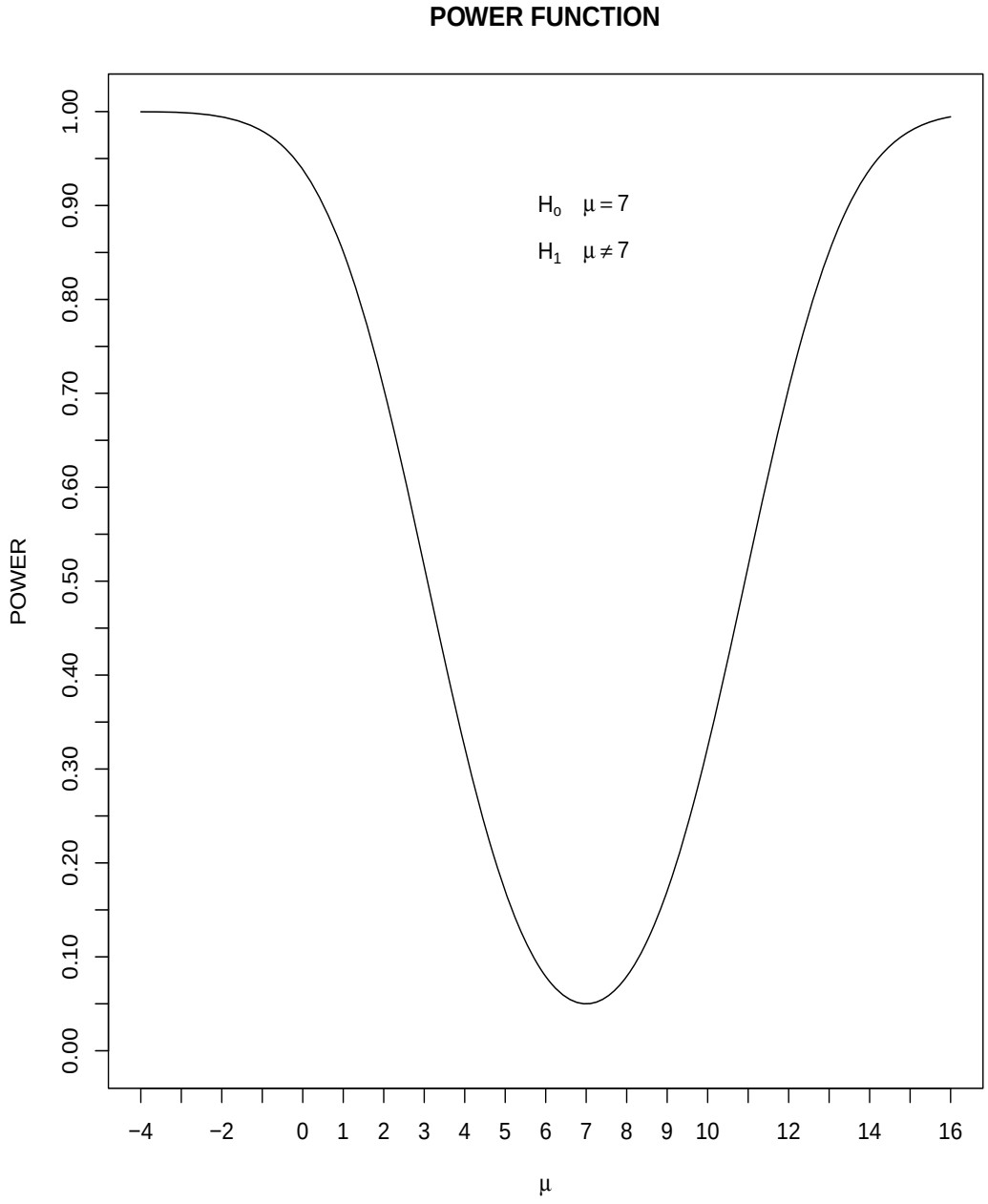
H2. For the one-sided hypotheses :
$$H_o : \theta \leq \theta_o \ \text{ versus } \ H_1 : \theta > \theta_o \qquad R = \{X : T(X) > C_{1-\alpha}\}$$

H3. For two-sided hypotheses:
$$H_o : \theta = \theta_o \ \text{ versus } \ H_1 : \theta \neq \theta_o \qquad R = \{X : T(X) < C_{\alpha/2} \quad \text{or} \quad T(X) > C_{1-\alpha/2}\}$$

For the above three sets of hypotheses, the power curves would have the following shapes:


POWER FUNCTION
POWER
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
H0 μ ≥ 8
H1 μ < 8
−2
−1
0
1
2
3
4
5
6
7
8
9
10
11
12
μ


POWER FUNCTION
H0 μ ≤ 2
H1 μ > 2
POWER
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
−2
0
2
4
6
8
10
12
14
μ


POWER FUNCTION
POWER
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
H0 μ = 7
H1 μ ≠ 7
−4
−2
0
1
2
3
4
5
6
7
8
9
10
12
14
16
μ

The following example will illustrate the how to determine the rejection region and compute the power curve.

Suppose we want to test the hypotheses:

$H_o : \mu \leq \mu_o$ versus $H_1 : \mu > \mu_o$

where μ is the mean of a population having a $N(\mu, \sigma^2)$ distribution.

Our null hypothesis has $\Theta_o = (-\infty, \mu_o]$ and

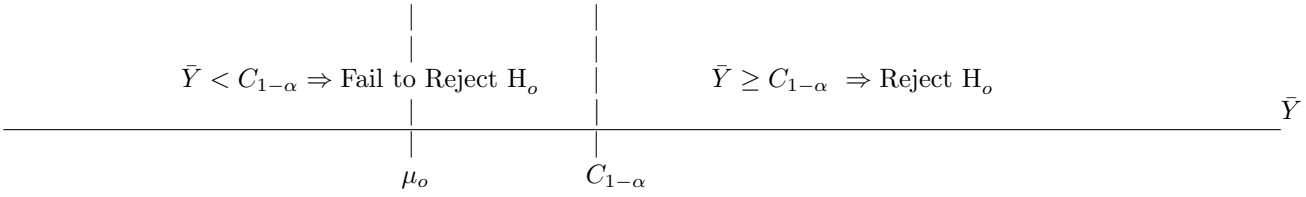
the alternative hypothesis has $\Theta_1 = (\mu_o, \infty)$.

Let $\mathbf{Y} = (Y_1, \cdots, Y_n)$ be a random sample from the population, i.e., $Y_1, \cdots, Y_n$ are iid $N(\mu, \sigma^2)$, we will consider the unrealistic situation where σ is known.

Our rejection region will be of the form: $R = \{\mathbf{Y} : T(\mathbf{Y}) > C_{1-\alpha}\} = \{\mathbf{Y} : \bar{Y} > C_{1-\alpha}\}$,

where our test statistic is $T(\mathbf{Y}) = \bar{Y}$ and $C_{1-\alpha}$ is the upper α-percentile of the sampling distribution of $T(\mathbf{Y})$.

Therefore, we will reject H_o when the sample mean $\bar{Y}$ is larger than μ_o, the question is how much larger, that is, how should we select $C_{1-\alpha}$?



The power function is then given by

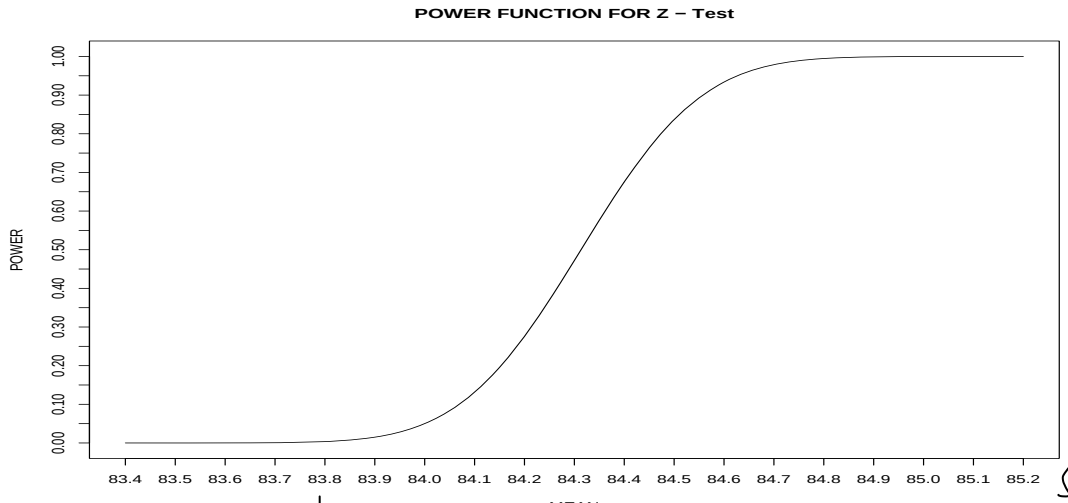
$$\gamma(\mu) = P_\mu\left(\text{reject} H_o\right) = P_\mu\left(\bar{Y} > C_{1-\alpha}\right) = P_\mu\left(\frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma} > \frac{\sqrt{n}(C_{1-\alpha}-\mu)}{\sigma}\right)$$

$$= P_\mu\left(Z > \frac{\sqrt{n}(C_{1-\alpha}-\mu)}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}(C_{1-\alpha}-\mu)}{\sigma}\right)$$

↳ above

$C_{1-\alpha}$ to maximize

the power



The max prob under the null of rejecting the null hypo has to be α. Bc the power function is monotonically increasing. That max prob happens at the boundary where $\mu = \mu_o$

The power function $\gamma(\mu)$ is increasing in μ because $\Phi\left(\frac{\sqrt{n}(C_{1-\alpha}-\mu)}{\sigma}\right)$ is decreasing in μ and hence 1- $\Phi\left(\frac{\sqrt{n}(C_{1-\alpha}-\mu)}{\sigma}\right)$ is increasing in μ. Therefore,

$$\sup_{\theta \epsilon \Theta_o} \gamma(\theta) = \sup_{\mu \leq \mu_o} \gamma(\mu) = \gamma(\mu_o) = 1 - \Phi\left(\frac{\sqrt{n}(C_{1-\alpha} - \mu_o)}{\sigma}\right).$$

Thus, if we want a significance level of α, we just need to set

$$1 - \Phi\left(\frac{\sqrt{n}(C_\alpha - \mu_o)}{\sigma}\right) = \alpha \Rightarrow \quad \frac{\sqrt{n}(C_{1-\alpha} - \mu_o)}{\sigma} = Z_\alpha \Rightarrow \quad C_{1-\alpha} = \mu_o + \frac{\sigma Z_\alpha}{\sqrt{n}}$$

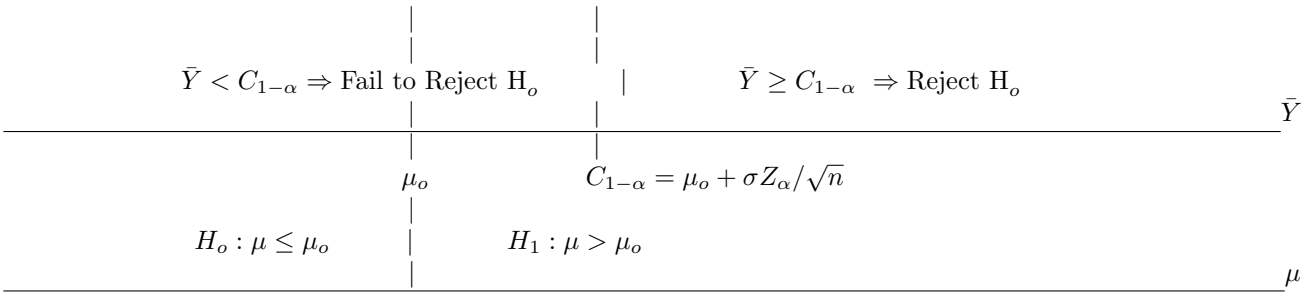
the power must = α

where $Z_\alpha = \Phi^{-1}(1-\alpha)$, is the $1-\alpha$ percentile or upper α percentile of the $N(0,1)$ distribution.

Thus, we reject H_o when

$$\frac{\sqrt{n}(\bar{Y} - \mu_o)}{\sigma} > Z_\alpha \quad \text{or} \quad \bar{Y} > \mu_o + \sigma Z_\alpha/\sqrt{n}$$

Based on the above rejection region, we have the following partition of the values of our point estimator of the population mean: $\hat{\mu} = \bar{Y}$ as a test statistic for testing $H_o : \mu \leq \mu_o$ vs $H_1 : \mu > \mu_o$.



When $\bar{Y}$ falls in the region (μ_o, C_α), $\bar{Y}$ is greater than μ_o but we still do not state that the data supports H_1. Why?

This region, (μ_o, C_α), is called the region of uncertainty. This region exists because of the Neyman-Pearson philosophy of testing: reject H_o only when there is **substantial evidence** in favor of H_1.

The size of this region is determined by how much certainty we demand of the data.

As we decrease α, our risk of making a Type I error, stating the data supports H_1, is decreased, C_α increases and we thus increase the size of our uncertainty region. We demand a greater amount of evidence in the data that H_1 is true.

Alternatively, our decision rule can be written as reject H_o when $Z = \frac{\sqrt{n}(\bar{Y}-\mu_o)}{\sigma} > Z_\alpha$.

This is the standard form of the decision rule for testing hypotheses about μ.

The power curves for various choices of n and α are given next. From these graphs we can observe the following relationships between sample size, the size of the test α, and the power of test.

1. For a fixed sample size, n, and value for the parameter being tested, μ, in this example,

   As the size of the test α increases from .001 to .05, the power increases, for example,

   For $\mu = 85$, $\gamma_{.001}(85) < \gamma_{.01}(85) < \gamma_{.05}(85)$

2. For a fixed sample size, n, and value for the parameter being tested, μ, in this example,

   As the size of the test α increases from .001 to .05, the probability of Type II error $\beta(\mu) = 1-\gamma(\mu)$ decreases, for example,

   For $\mu = 85$, $\beta_{.001}(85) > \beta_{.01}(85) > \beta_{.05}(85)$

3. For a fixed size of the test, α, and value for the parameter being tested, μ, in this example,

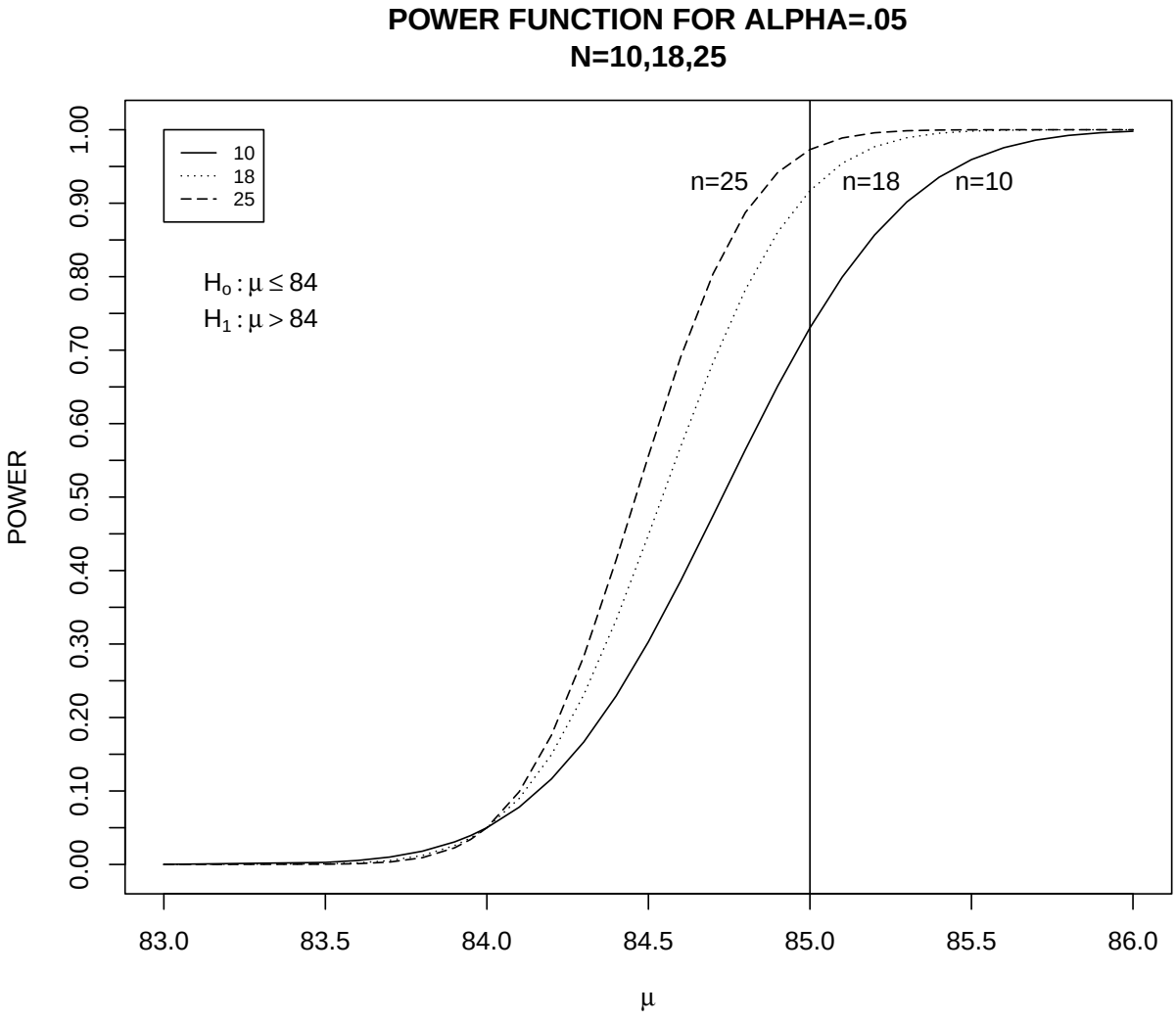
   As the sample size n increases from 10 to 25, the power increases, for example,

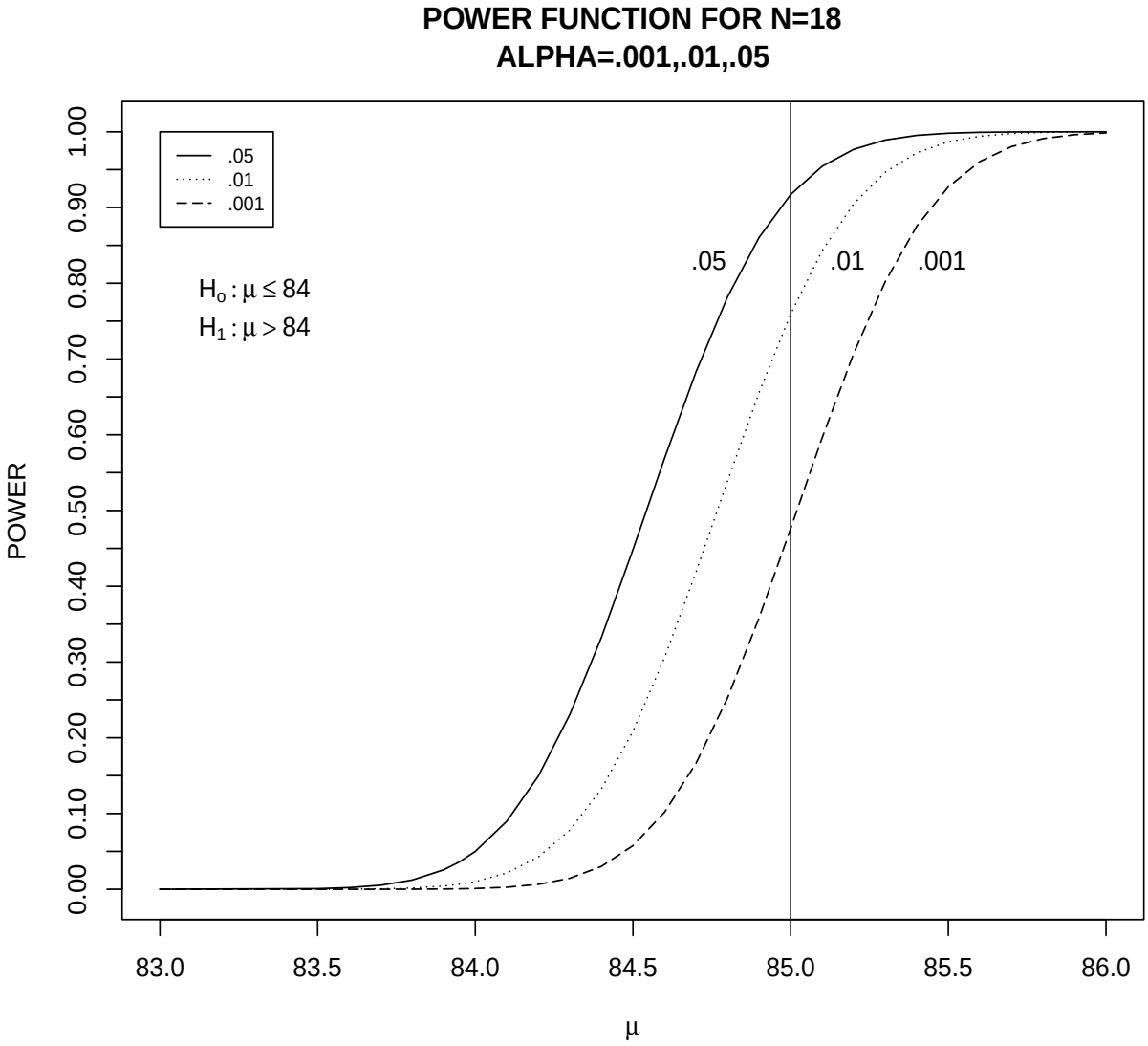
   For $\mu = 85$, $\gamma_{10}(85) < \gamma_{18}(85) < \gamma_{25}(85)$

   As the sample size n increases from 10 to 25, the probability of Type II error $\beta(\mu) = 1 - \gamma(\mu)$ decreases, for example,

   For $\mu = 85$, $\beta_{10}(85) > \beta_{18}(85) > \beta_{25}(85)$

   For fixed α and θ, $\beta_n(\theta)$ decreases as n increases and $\gamma_n(\theta)$ increases as n increases


POWER FUNCTION FOR ALPHA=.05
N=10,18,25
POWER
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
10
18
25
$H_0 : \mu \leq 84$
$H_1 : \mu > 84$
n=25
n=18
n=10
83.0
83.5
84.0
84.5
85.0
85.5
86.0
μ


POWER FUNCTION FOR N=18
ALPHA=.001,.01,.05
.05
.01
.001
$H_0 : \mu \leq 84$
$H_1 : \mu > 84$
.05
.01
.001
POWER
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
83.0
83.5
84.0
84.5
85.0
85.5
86.0
μ

Summary of the Basics of Hypothesis Testing.

1. From the nature of the experimental setting, identify the appropriate probability model and translate each statement about the population or process in terms of the range of values for the relevant model parameters.

2. When evidence is sought to establish a particular assertion about the population or process using a random sample from the population or process, the assertion is formulated as the research hypothesis H_1 and the negation of the assertion is formulated as the null hypothesis H_o.

3. Select a function of the data and the parameter being tested, called the test statistic, whose value is a measure of the plausibility of the research hypothesis. The test statistic is often formulated using the theory of UMP tests or using likelihood ratio principles.

4. Specify the designed level of significance, α, the maximum probability of Type I error. Using the sampling distribution of the test statistic, determine a rejection region for the test such that the maximum probability of Type I error does not exceed α. Generally, it is sufficient to determine the rejection region such that the probability of Type I error at the boundary between H_o and H_1 is equal to α. Define the power of the test to be the probability that the test statistic rejects H_o as a function of the values of the parameter being tested: The **power function** is given by

$$\gamma(\theta^*) = Pr[\text{Test Statistic is in Rejection Region when value of parameter } \theta = \theta^*]$$

   Power function is defined over the whole parameter space, Θ, not just the alternative region, Θ_1.

5. The power curve of the test should now be examined to ensure that the choice of sample size and α provides reasonable protection against Type II errors at crucial values of the parameter in H_1. If the probability of Type II error is too high, then the sample size must be increased and/or the rejection region adjusted by increasing α in order to have a proper balance between the probabilities of Type I and Type II errors.

6. After the test procedure has been explicitly formulated, compute the value of the test statistic from the data and determine whether or not the test statistic falls in the rejection region. If it does, conclude that the data supports the research hypothesis, H_1 at the specified level of α. That is, there is a probability of size α that a Type I error has been committed. If the computed value of the test statistic falls outside the rejection region, then state that the data does not support the research hypothesis and provide a power curve to inform the researcher about the possibilities of a Type II error. In neither case, do we state that the validity of H_o or H_1 has been **PROVEN** by the data. There is always the possibility that a Type I or a Type II error has been committed by the decision rule. We attempt to control the probability of Type I and Type II errors by our selection of the test statistics and sample size, n.

   The decision to reject or fail to reject H_o is based on which of H_o and H_1 is more plausible based on the observed data. We are not proving that one of the hypotheses, H_o or H_1, is true but simply indicating which of the two hypotheses is more strongly supported by the data.

7. Alternative Approach: Define **p-value**, called the **significance probability**, as the smallest value of α which would lead to the rejection of H_o.

   • p-value is the probability of a obtaining a value of the test statistic as extreme or more extreme to H_o than the value computed from the observed data

   • p-value is the probability to obtain a value of the test statistic which is as likely or more likely to reject H_o.

   A decision rule is then formulated using the p-value:

   If p-value$\leq \alpha$, then conclude H_1 is supported by data, i.e., Reject H_o

   If p-value$> \alpha$, then conclude H_1 is not supported by data, i.e., Fail to Reject H_o

   • p-value is **not** the probability the H_o is true.

   • p-value is the probability of obtaining a new data set under the conditions specified in H_o, such that the new data set yields greater evidence that H_1 is true than is indicated using the current data. The p-value is computed using the same test statistics as was used in the original data.

   • Prior to conducting the experiment, the researcher **must** select a value of α to reflect his/her assessment of the risk of a Type I error. This value is the desired level of significance of the testing procedure. This approach provides a decision between H_o and H_1 along with a measure of the strength of the decision, the p-value.

   • Some researchers do not use the above decision rule. They just report the value of the p-value and let the reader of the research paper make their own assessment of whether or not the research hypothesis has been supported by the data.

8. Finally, whenever possible, provide a confidence interval for the parameter so that the researchers can assess the practical significance of their decision (as opposed to the statistical significance).

   Estimation of the effect size.

9. The American Statistical Association has issued a statement about statistical significance and p-values.

ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®
732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

Good statistical practice is an essential component of good scientific practice, the statement observes, and such practice "emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean."

"The *p*-value was never intended to be a substitute for scientific reasoning," said Ron Wasserstein, the ASA's executive director. "Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post $p<0.05$ era.'"

"Over time it appears the *p*-value has become a gatekeeper for whether work is publishable, at least in some fields," said Jessica Utts, ASA president. "This apparent editorial bias leads to the 'file-drawer effect,' in which research with statistically significant outcomes are much more likely to get published, while other work that might well be just as important scientifically is never seen in print. It also leads to practices called by such names as '*p*-hacking' and 'data dredging' that emphasize the search for small *p*-values over other statistical and scientific reasoning."

The statement's six principles, many of which address misconceptions and misuse of the *p*-value, are the following:

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

The statement has short paragraphs elaborating on each principle.

In light of misuses of and misconceptions concerning *p*-values, the statement notes that statisticians often supplement or even replace *p*-values with other approaches. These include methods "that emphasize estimation over testing such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence such as likelihood ratios or Bayes factors; and other approaches such as decision-theoretic modeling and false discovery rates."

"The contents of the ASA statement and the reasoning behind it are not new—statisticians and other scientists have been writing on the topic for decades," Utts said. "But this is the first time that the community of statisticians, as represented by the ASA Board of Directors, has issued a statement to address these issues."

"The issues involved in statistical inference are difficult because inference itself is challenging," Wasserstein said. He noted that more than a dozen discussion papers are being published in the ASA journal *The American Statistician* with the statement to provide more perspective on this broad and complex topic. "What we hope will follow is a broad discussion across the scientific community that leads to a more nuanced approach to interpreting, communicating, and using the results of statistical methods in research."

About the American Statistical Association

The ASA is the world's largest community of statisticians and the oldest continuously operating professional science society in the United States. Its members serve in industry, government and academia in more than 90 countries, advancing research and promoting sound statistical

START CN Friday 11/5/2021

Tests About Population/Process Mean μ

Case 1: Normal Distribution with σ known

Suppose $Y_1, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$ distributed with either n large or σ known. A test statistic for testing hypotheses about μ ($H_1 : \mu < \mu_o$, $H_1 : \mu > \mu_o$, or $H_1 : \mu \neq \mu_o$ with μ_o specified) is

$$Z = \frac{\sqrt{n}(\bar{Y} - \mu_o)}{\sigma}$$

In the following let $\bar{y}$ be the observed value of $\bar{Y}$

The distribution of Z is $N\left(\frac{\sqrt{n}(\mu-\mu_o)}{\sigma}, 1\right)$ and hence depends on the value of μ.

The form of the power function and p-value for the test depend on H_1 and are given in terms of the N(0,1) cdf Φ which can be obtained from R through the expression: $\Phi(z) = pnorm(z)$

1. With $H_o : \mu \leq \mu_o$ vs $H_1 : \mu > \mu_o$:

   Reject H_o if $\bar{Y} > C_\alpha = \mu_o + z_\alpha \frac{\sigma}{\sqrt{n}}$, where z_α is upper α percentile from the N(0,1) distribution.

   Power Function:

$$\gamma(\mu) = P[\text{Reject } H_o] = P_\mu\left(\bar{Y} > \mu_o + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = P_\mu\left(\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} > \frac{\mu_o-\mu}{\sigma/\sqrt{n}} + z_\alpha\right) = 1 - \Phi\left(z_\alpha + \frac{\sqrt{n}(\mu_o-\mu)}{\sigma}\right)$$

$$p - value = P_{\mu_o}\left(Z > \frac{\sqrt{n}(\bar{y} - \mu_o)}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}(\bar{y} - \mu_o)}{\sigma}\right)$$

2. With $H_o : \mu \geq \mu_o$ vs $H_1 : \mu < \mu_o$:

   Reject H_o if $\bar{Y} < \mu_o - z_\alpha \frac{\sigma}{\sqrt{n}}$

$$\text{Power Function: } \gamma(\mu) = P_\mu\left(\bar{Y} < \mu_o - z_\alpha \frac{\sigma}{\sqrt{n}}\right) = \Phi\left(-z_\alpha + \frac{\sqrt{n}(\mu_o - \mu)}{\sigma}\right)$$

$$p - value = P_{\mu_o}\left(Z < \frac{\sqrt{n}(\bar{y} - \mu_o)}{\sigma}\right) = \Phi\left(\frac{\sqrt{n}(\bar{y} - \mu_o)}{\sigma}\right)$$

3. With $H_o : \mu = \mu_o$ vs $H_1 : \mu \neq \mu_o$:

   Reject H_o if

$\bar{Y} < \mu_o - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or $\bar{Y} > \mu_o + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, that is, $\left|\frac{\sqrt{n}(\bar{Y}-\mu_o)}{\sigma}\right| > z_{\alpha/2}$

$$\text{Power Function: } \gamma(\mu) = P_\mu\left(\left|\frac{\sqrt{n}(\bar{Y} - \mu_o)}{\sigma}\right| > z_{\alpha/2}\right) = \Phi\left(-z_{\alpha/2} + \frac{\sqrt{n}(\mu_o - \mu)}{\sigma}\right) + 1 - \Phi\left(z_{\alpha/2} + \frac{\sqrt{n}(\mu_o - \mu)}{\sigma}\right)$$

$$p - value = P_{\mu_o}\left(|Z| > \left|\frac{\sqrt{n}(\bar{y} - \mu_o)}{\sigma}\right|\right) = 2\left(1 - \Phi\left(\left|\frac{\sqrt{n}(\bar{y} - \mu_o)}{\sigma}\right|\right)\right)$$

Example The problem is to test $H_o : \mu \leq 30$ versus $H_1 : \mu > 30$ in a $N(\mu, 81)$ populations using a sample of size n=25.

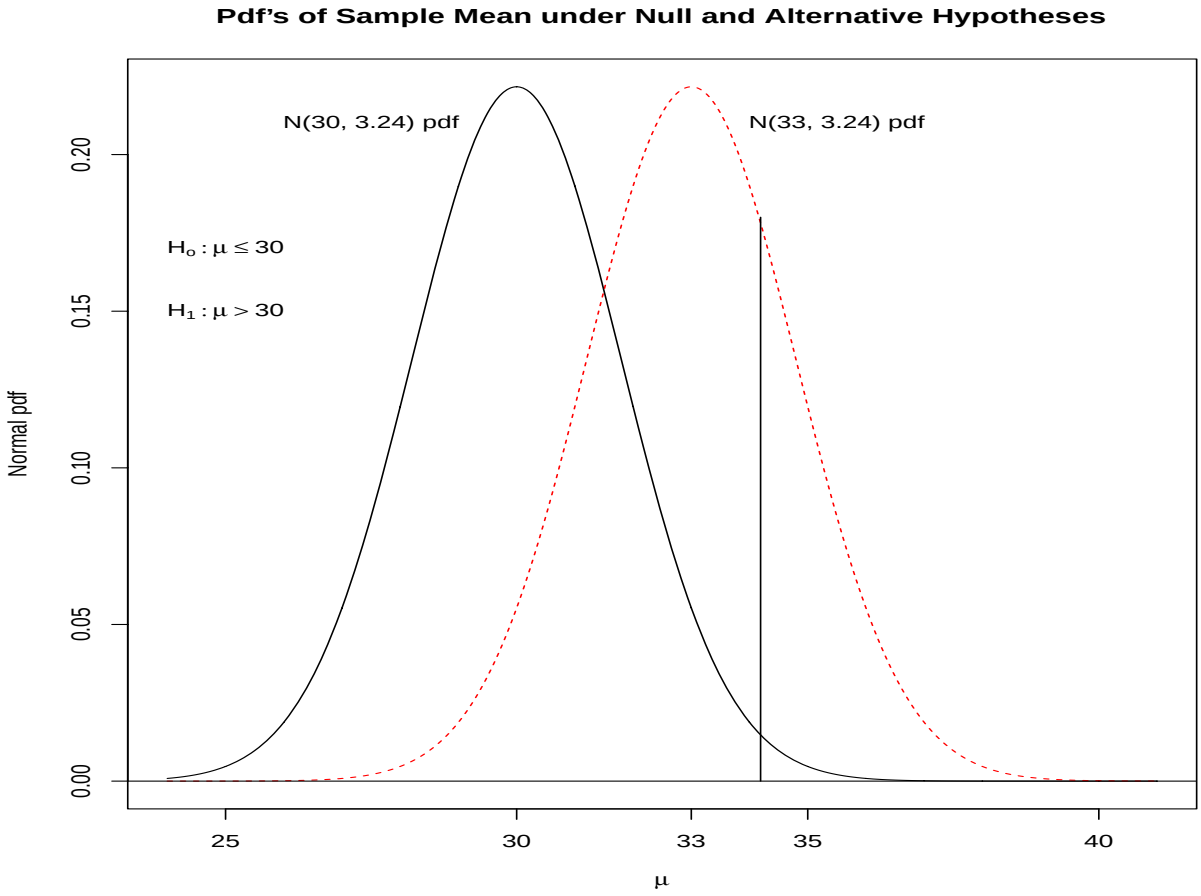
With a test of size $\alpha = .01$, reject H_o if $\bar{Y} \geq 30 + Z_{.01}\sigma/\sqrt{n} = 30 + (2.326)(9)/\sqrt{25} = 34.19$.

When $\mu = 30$, the sampling distribution of $\bar{Y}$ is $N(30, 81/25) = N(30, 3.24)$

When $\mu = 33$, the sampling distribution of $\bar{Y}$ is $N(33, 81/25) = N(33, 3.24)$

The area under the $N(30, 3.24)$ beyond 34.19 is $\gamma(30) = .01$ (the value of α)

The area under the $N(33, 3.24)$ beyond 34.19 is $\gamma(33) = .2543$ (power at $\mu = 33$).



Sample Size Determination

Determine the sample size n such that a size α test used to evaluate the data from the experiment (study) will have sufficient power to detect a deviation from μ_o that is of practical significance to the researcher.

1. Let δ be the size of the deviation from μ_o that the researcher wants to detect.
2. Specify the size α of the test
3. Specify the region that is of practical significance:

   the population mean μ is at least δ units from μ_o
4. Specify the probability of detecting this region, that is,

   The power should be at least $1 - \beta$ to detect this region, or

   The probability of Type II error is at most β of failing to detect the region

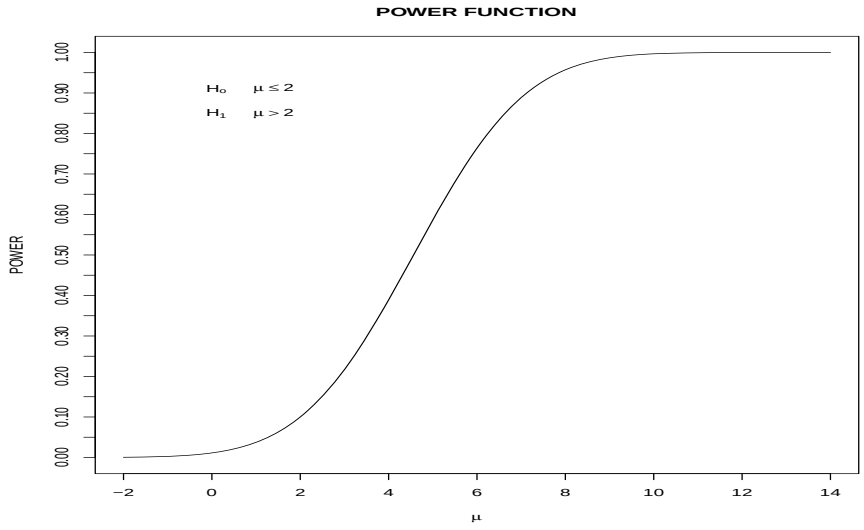
We need to find the minimum sample size n to meet these specifications.

That is, find the minimum n such that

$$\gamma(\mu) \geq 1 - \beta \quad \text{whenever} \quad |\mu - \mu_o| \geq \delta \quad \text{for} \quad \mu \in \Theta_1 \quad \text{or equivalently}$$

$$\beta(\mu) = 1 - \gamma(\mu) \leq \beta \quad \text{whenever} \quad |\mu - \mu_o| \geq \delta \quad \text{for} \quad \mu \in \Theta_1$$

We will first consider the case $H_1 : \mu > \mu_o$.



Because the power function $\gamma(\mu)$ is increasing in μ, we only need to find n to satisfy the bound at $\mu_1 = \mu_o + \delta$. For all other values of μ_1 greater than $\mu_o + \delta$ the power will be larger than the power at $\mu_o + \delta$.

$$\begin{aligned}
\gamma(\mu_o+\delta) &= P[\bar{X} > \mu_o + z_\alpha \frac{\sigma}{\sqrt{n}} \text{ when } \mu = \mu_o + \delta] \\
&= P\left[\frac{\bar{X}-(\mu_o+\delta)}{\sigma/\sqrt{n}} > z_\alpha + \frac{-\delta\sqrt{n}}{\sigma}\right] \\
&= 1 - \Phi\left(z_\alpha + \frac{-\delta\sqrt{n}}{\sigma}\right) = 1-\beta
\end{aligned}$$

With $\Phi(z_\beta) = 1-\beta \Rightarrow \Phi(-z_\beta) = \beta$, we have

$$-z_\beta = z_\alpha + \frac{-\delta\sqrt{n}}{\sigma} \quad \Rightarrow \quad n = \left[\frac{\sigma(z_\alpha + z_\beta)}{\delta}\right]^2$$

We obtain the same formula for n in the case of $H_o : \mu \geq \mu_o$ vs $H_1 : \mu < \mu_o$.

EXAMPLE Suppose the researcher wants to test $H_o : \mu \leq 20$ vs $H_1 : \mu > 20$. Based on past studies, she states that the population has a normal distribution with $\sigma \approx 40$.

She wants to determine the value of n such that a size $\alpha = .05$ test will have power of at least 90% of detecting that the population mean is at least 35.

Thus, we have the following specifications

$\delta = 35 - 20 = 15$

$\alpha = .05$

$\beta = 1 - .9 = .1$

$\sigma = 40$

Therefore, the minimum sample size needed to meet these requirements is

$$\begin{aligned}
n &= \left[\frac{\sigma(z_\alpha + z_\beta)}{\delta}\right]^2 \\
&= \left[\frac{40(1.645+1.282)}{15}\right]^2 \\
&= 60.9 \quad \Rightarrow \\
& \\
& n \geq 61
\end{aligned}$$

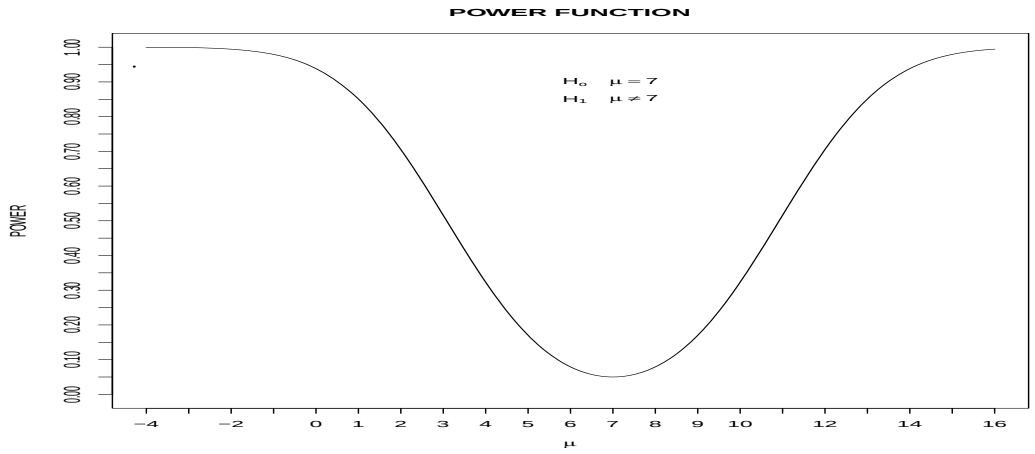
Two-Sided Alternative

Goal: Find the sample size n such that a size α test will have power at μ greater than $1-\beta_o$ whenever $|\mu - \mu_o| \geq \delta$

For the two-sided hypothesis, $H_o : \mu = \mu_o$ vs $H_1 : \mu \neq \mu_o$, we have that the power function is symmetric about μ_o and increasing in $\delta = |\mu - \mu_o|$.

Thus, we have the following specifications:

find the minimum n such that $\gamma(\mu_o + \delta) = 1 - \beta$ or $\gamma(\mu_o - \delta) = 1 - \beta$.



Thus, find the minimum n to satisfy $\gamma(\mu_o + \delta) = 1 - \beta$:

$$\gamma(\mu_o + \delta) = \Phi\left(-z_{\alpha/2} + \frac{-\delta\sqrt{n}}{\sigma}\right) + 1 - \Phi\left(z_{\alpha/2} + \frac{-\delta\sqrt{n}}{\sigma}\right) = 1 - \beta$$

There is no closed form solution in n for the above equation.

However, for reasonable values of σ and δ, $\Phi\left(-z_{\alpha/2} + \frac{-\delta\sqrt{n}}{\sigma}\right)$ is negligible relative to the other term in the equation.

Thus, an approximate solution can be obtained from

$$\gamma(\mu_o + \delta) \approx 1 - \Phi\left(z_{\alpha/2} + \frac{-\delta\sqrt{n}}{\sigma}\right) = 1 - \beta$$

Thus, we have with $\Phi(z_\beta) = 1 - \beta \Rightarrow \Phi(-z_\beta) = \beta$,

$$-z_\beta = z_{\alpha/2} + \frac{-\delta\sqrt{n}}{\sigma} \quad \Rightarrow \quad n = \left[\frac{\sigma(z_{\alpha/2} + z_\beta)}{\delta}\right]^2$$

The above formulas require knowledge of σ.

The researcher will need to provide this value from previous studies, pilot studies, or the literature.

The most difficult part of these calculations is obtaining from the researcher what values they want for α, δ, $1-\beta$, and their guess at σ

Case 2: Normal Population with σ unknown

Suppose $Y_1, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$ distributed with σ unknown. A test statistic for testing hypotheses about μ ($H_1 : \mu < \mu_o$, $H_1 : \mu > \mu_o$, or $H_1 : \mu \neq \mu_o$ with μ_o specified) is

$$T = \frac{\sqrt{n}(\bar{Y} - \mu_o)}{S}$$

replace σ w/ sample SD S

We can rewrite T as

$$T = \frac{\frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma} + \frac{\sqrt{n}(\mu-\mu_o)}{\sigma}}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}} \text{ distributed as } \frac{N(0,1) + \Delta}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

Thus, the distribution of T is a non-central $t-$distribution with $df = n - 1$ and

noncentrality parameter $\Delta = \frac{\sqrt{n}(\mu-\mu_o)}{\sigma}$ which depends on the value of μ and σ.

This would seem to cause great difficulty in setting up tests of hypotheses and computing p-values.

However, both of these probability calculations are conducted when $\mu = \mu_o$ and hence $\Delta = 0$. Thus, we can use the central t-distribution which just depends on the sample size.

The form of the power function and p-value for the test depend on the specification of H_1:

1. With $H_o : \mu \leq \mu_o$ vs $H_1 : \mu > \mu_o$ and t_α the upper α percentile from a central t-distribution with $df = n - 1$,

   Reject H_o if $T > t_\alpha = qt(1 - \alpha, n - 1)$

   Power Function: $\gamma(\mu) = P_\mu(T > t_\alpha) = 1 - G(t_\alpha) = 1 - pt(qt(1 - \alpha, n - 1), n - 1, \Delta)$

   where G is the cdf of a noncentral t-distribution with $df = n - 1$ and noncentrality parameter and $\Delta = \frac{\sqrt{n}(\mu-\mu_o)}{\sigma}$.

   $$p - value = P_{\mu_o}\left(T > \frac{\sqrt{n}(\bar{Y} - \mu_o)}{S}\right) = 1 - G_o\left(\frac{\sqrt{n}(\bar{Y} - \mu_o)}{S}\right) = 1 - pt\left(\frac{\sqrt{n}(\bar{Y} - \mu_o)}{S}, n - 1\right)$$

   where G_o is the cdf of a central t-distribution with $df = n - 1$.

2. With $H_o : \mu \geq \mu_o$ vs $H_1 : \mu < \mu_o$ and t_α the upper α percentile from a central t-distribution with $df = n - 1$,

   Reject H_o if $T < -t_\alpha = qt(\alpha, n - 1) \Rightarrow \bar{Y} < \mu_o - t_\alpha S/\sqrt{n}$

   Power Function: $\gamma(\mu) = P_\mu(T < -t_\alpha) = G(-t_\alpha) = pt(-qt(1 - \alpha, n - 1), n - 1, \Delta)$

   where G is the cdf of a noncentral t-distribution with $df = n - 1$ and noncentrality parameter $\Delta = \frac{\sqrt{n}(\mu-\mu_o)}{\sigma}$

   $$p - value = P_{\mu_o}\left(T < \frac{\sqrt{n}(\bar{Y} - \mu_o)}{S}\right) = G_o\left(\frac{\sqrt{n}(\bar{Y} - \mu_o)}{S}\right) = pt\left(\frac{\sqrt{n}(\bar{Y} - \mu_o)}{S}, n - 1\right)$$

   where G_o is the cdf of a central t-distribution with $df = n - 1$.

3. With $H_o : \mu = \mu_o$ vs $H_1 : \mu \neq \mu_o$ and $t_{\alpha/2}$ the upper $\alpha/2$ percentile from a central t-distribution with $df = n - 1$,

   Reject H_o if $|T| > t_{\alpha/2}$

$$\text{Power Function: } \gamma(\mu) = P_\mu\left(|T| > t_{\alpha/2}\right) = G(-t_{\alpha/2}) + 1 - G(t_{\alpha/2}),$$

   where G is the cdf of a noncentral t-distribution with $df = n - 1$ and noncentrality parameter $\Delta = \frac{\sqrt{n}(\mu - \mu_o)}{\sigma}$

$$\begin{aligned} p - value &= P_{\mu_o}\left(|T| > \left|\frac{\sqrt{n}(\bar{Y} - \mu_o)}{S}\right|\right) \\ &= 2P_{\mu_o}\left(T > \left|\frac{\sqrt{n}(\bar{Y} - \mu_o)}{S}\right|\right) \\ &= 2\left(1 - G_o\left(\left|\frac{\sqrt{n}(\bar{Y} - \mu_o)}{S}\right|\right)\right), \end{aligned}$$

   where G_o is the cdf of a central t-distribution with $df = n - 1$, hence

$$p - value = 2\left(1 - pt\left(\left|\frac{\sqrt{n}(\bar{Y} - \mu_o)}{S}\right|, n - 1\right)\right)$$

EXAMPLE Suppose the researcher wants a size $\alpha = .05$ test of $H_o : \mu \leq 84$ vs $H_1 : \mu > 84$ using $n = 10$ observations from a normal population with mean μ and unknown σ.

The test would reject H_o when $T = \frac{\sqrt{10}(\bar{Y}-84)}{S} > t_{.05,9} = 1.833 = qt(1 - .05, 9)$.

From the data, $\bar{Y} = 86.3$ and $S = 3.2$, thus

$$T = \frac{\sqrt{10}(86.3 - 84)}{3.2} = 2.27 > 1.833 \quad \Rightarrow \quad \text{Reject} \quad H_o$$

and thus conclude that the data supports the statement: $H_1 : \mu > 84$

A 95% lower bound on μ is given by

$$\bar{Y} \; - \; t_{.05,9} \, \frac{S}{\sqrt{n}} \; = \; 86.3 \; - \; 1.833 \, \frac{3.2}{\sqrt{10}} \; = \; (84.45, \; \infty)$$

Note that the lower bound on μ is greater than $\mu_o = 84$ which supports the contention that $\mu > 84$

To further reflect the strength of our conclusion, we use the central t cdf G with $df = 9$ to compute

$$p - value = P_{\mu_o}(T > 2.27) = 1 - G_o(2.27) = 1 - pt(2.27, 9) = .0247 < .05 = \alpha$$

Using the R function, $pt(2.27, 9)$

Thus, we would have rejected H_o for any $\alpha \geq .0247$.

To compute the power for this test we need to use the R function for the cdf of a

noncentral t-distribution with df = 9 and noncentrality parameter

$\Delta = \frac{\sqrt{10}(\mu-84)}{\sigma}$: $\quad \gamma(\mu) = 1 - pt(qt(.95, df), df, \Delta)$,

where $qt(.95, df)$ is the upper $\alpha = .05$ percentile of a central t-distribution with $df = n - 1$,

Δ would be a vector of values for the noncentrality parameter, using a given value of σ, we will take $\sigma = 1.4$ for the purpose of this example.

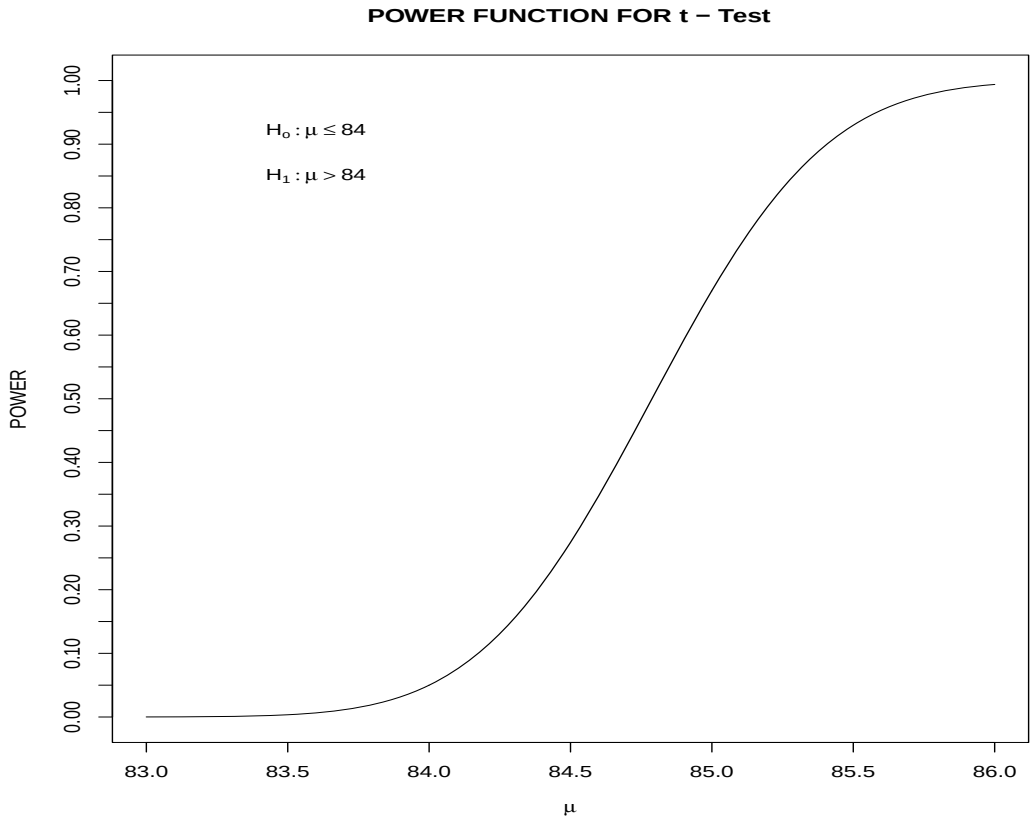
pt is the cdf of a noncentral t-distribution:

μ	83.0	83.5	83.7	83.9	83.95	84	84.2	84.4
Δ	-2.26	-1.13	-0.68	-0.23	-0.11	0.00	0.45	0.90
$\gamma(\mu)$ or $\gamma(\Delta)$	.00009	.0035	.0115	.032	.040	.05	.110	.209
μ	84.6	84.8	85	85.2	85.4	85.6	85.8	86
Δ	1.36	1.81	2.26	2.71	3.16	3.61	4.07	4.52
$\gamma(\mu)$ or $\gamma(\Delta)$	.348	.510	.670	.803	.900	.950	.982	.993

A plot of the power curve is given on the next page:

R code:

```
n =  10
df =  n-1
muo  =  84
sigma =  1.4
mu =   seq(83,86,.05)
delta =  sqrt(n)*(mu-muo)/sigma
power =  1-pt(qt(.95,df),df,delta)
par(lab=c(15,20,4))
plot(mu,power,type="l",ylim=c(0,1),xlab=expression(mu),
ylab="POWER")
title("POWER FUNCTION FOR t - Test")
out = cbind(mu,delta,power)
```



Selection of H_o and H_1 ← minor point.

Case 1: Test $H_o : \mu \leq 84$ vs $H_1 : \mu > 84$ at the $\alpha = .01$ level.

Data: n=10, $\bar{Y} = 86.3$, $S = 3.2$

Reject H_o if $\bar{Y} \geq 84 + 2.821\frac{3.2}{\sqrt{10}} = 86.85$, where $t_{.01,9} = 2.821$

$\bar{Y} = 86.3 < 86.85$ therefore, we fail to reject H_o and conclude the evidence in the data does not support $H_1 : \mu > 84$

Does the selection of the form of H_1 alter the conclusions that we obtain in hypotheses testing?

Suppose the hypotheses in this problem were erroneously set up as $H_o : \mu \geq 84$ vs $H_1 : \mu < 84$.

Would we still have the same conclusion as with the other arrangement?

That is, would we still conclude that the data did not support the statement $H_1 : \mu > 84$.

Case 2: $H_o : \mu \geq 84$ vs $H_1 : \mu < 84$, Reject H_o if $T = \frac{\sqrt{10}(\bar{Y}-84)}{S} < -t_{.01,9} = 2.821$

$T = \frac{\sqrt{10}(86.3-84)}{3.2} = 2.27 > -2.821 \quad \Rightarrow \quad$ Fail to Reject H_o

and thus conclude that the data does not support the statement: $H_1 : \mu < 84$

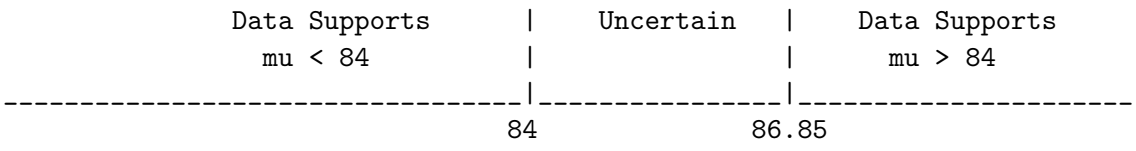
In the arrangement $H_o : \mu \geq 84$ vs $H_1 : \mu < 84$, we stated the data tends to support $\mu \geq 84$

In the arrangement $H_o : \mu \leq 84$ vs $H_1 : \mu > 84$, we stated the data tends to support $\mu \leq 84$

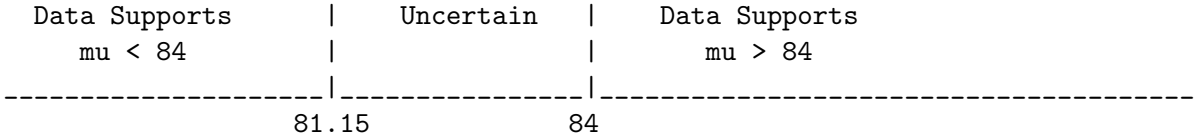
Why does a contradiction in conclusions occur?

The region of uncertainty has been shifted from values larger than 84 to values smaller than 84 in the two arrangements of H_o.

Case 1: Test $H_o : \mu \leq 84$ vs $H_1 : \mu > 84$



Case 2: Test $H_o : \mu \geq 84$ vs $H_1 : \mu < 84$



Sample Size Determination

The problem is to find the sample size n for a size α test such that the power at μ is greater than $1-\beta$, that is,

$\gamma(\mu) \geq 1-\beta$, whenever $|\mu - \mu_o| \geq \delta$.

The solution is much more difficult to obtain in this case because the power function depends on the degrees of freedom, $df = n-1$ which also depends on the sample size, n, the quantity we are attempting to find.

Thus, we must solve the equations for the power function iteratively to obtain the required sample sizes.

There are tables for determining the required sample sizes when using the t-distribution.

Table A11 from *Statistical Design and Analysis of Experiments*, by R. Mason, R. Gunst, J. Hess is given on the next page.

In this table, we need to specify

1. the size of the test α,
2. maximum probability of Type II error β,
3. size of difference to be detected: $\phi = \frac{|\mu-\mu_o|}{\sigma}$,

This table provides the minimum sample size n to have $\gamma(\mu) \geq 1-\beta$, that is,

the minimum sample size n to have $\beta(\mu) \leq \beta$ whenever, $|\mu - \mu_o| \geq \delta$.

For example, suppose we want to design a study to test $H_o : \mu \leq 84$ vs $H_1 : \mu > 84$.

The researcher wants the size of the test to be .05 and wants the test to have power of at least 80% to detect that μ is greater 84.8.

From previous studies, she is fairly certain that $\sigma < 2$.

Find the minimum n to achieve this specifications.

1. $\alpha = .05$
2. $\beta = 1 - .8 = .2$
3. $\phi = \frac{|84.8-84|}{2} = .4$.

   From the table read $n = 40$.

Using the R function,

```
power.t.test(n=,delta=.8,sd=2,sig.level=.05,power=.8,type=c("one.sample"),alternative=c("one.sided"))
```

We obtain, n = 40.029 with power using n=40 given by

```
power.t.test(n=40,delta=.8,sd=2,sig.level=.05,power=,type=c("one.sample"),alternative=c("one.sided"))
```

Power = 0.7997 Thus, to be exact you would need to have n=41 but in most cases n=40 would be acceptable.

Table A11 Sample-Size Requirements for Tests on the Mean of a Normal Distribution, Unknown Standard Deviation

		Level of t Test												
Single-Sided Test		$\alpha = 0.005$				$\alpha = 0.01$				$\alpha = 0.025$				$\alpha = 0.05$				
Double-Sided Test		$\alpha = 0.01$				$\alpha = 0.02$				$\alpha = 0.05$				$\alpha = 0.1$				
β =		0.01	0.05	0.1	0.2	0.01	0.05	0.1	0.2	0.01	0.05	0.1	0.2	0.01	0.05	0.1	0.2	
	0.05																	0.05
	0.10																	0.10
	0.15																	0.15
	0.20																	0.20
	0.25												128			139	101	0.25
	0.30				134				115			119	90		122	97	71	0.30
	0.35			125	99			109	85		109	88	67		90	72	52	0.35
	0.40		115	97	77		101	85	66	117	84	68	51	101	70	55	40	0.40
	0.45		92	77	62	110	81	68	53	93	67	54	41	80	55	44	33	0.45
	0.50	100	75	63	51	90	66	55	43	76	54	44	34	65	45	36	27	0.50
	0.55	83	63	53	42	75	55	46	36	63	45	37	28	54	38	30	22	0.55
	0.60	71	53	45	36	63	47	39	31	53	38	32	24	46	32	26	19	0.60
	0.65	61	46	39	31	55	41	34	27	46	33	27	21	39	28	22	17	0.65
	0.70	53	40	34	28	47	35	30	24	40	29	24	19	34	24	19	15	0.70
	0.75	47	36	30	25	42	31	27	21	35	26	21	16	30	21	17	13	0.75
	0.80	41	32	27	22	37	28	24	19	31	22	19	15	27	19	15	12	0.80
Value of	0.85	37	29	24	20	33	25	21	17	28	21	17	13	24	17	14	11	0.85
$\phi = \frac{\lvert\delta\rvert}{\sigma}$	0.90	34	26	22	18	29	23	19	16	25	19	16	12	21	15	13	10	0.90
	0.95	31	24	20	17	27	21	18	14	23	17	14	11	19	14	11	9	0.95
	1.00	28	22	19	16	25	19	16	13	21	16	13	10	18	13	11	8	1.00
	1.1	24	19	16	14	21	16	14	12	18	13	11	9	15	11	9	7	1.1
	1.2	21	16	14	12	18	14	12	10	15	12	10	8	13	10	8	6	1.2
	1.3	18	15	13	11	16	13	11	9	14	10	9	7	11	8	7	6	1.3
	1.4	16	13	12	10	14	11	10	9	12	9	8	7	10	8	7	5	1.4
	1.5	15	12	11	9	13	10	9	8	11	8	7	6	9	7	6		1.5
	1.6	13	11	10	8	12	10	9	7	10	8	7	6	8	6	6		1.6
	1.7	12	10	9	8	11	9	8	7	9	7	6	5	8	6	5		1.7
	1.8	12	10	9	8	10	8	7	7	8	7	6		7	6			1.8
	1.9	11	9	8	7	10	8	7	6	8	6	6		7	5			1.9
	2.0	10	8	8	7	9	7	7	6	7	6	5		6				2.0
	2.1	10	8	7	7	8	7	6	6	7	6			6				2.1
	2.2	9	8	7	6	8	7	6	5	7	6			6				2.2
	2.3	9	7	7	6	8	6	6		6	5			5				2.3
	2.4	8	7	7	6	7	6	6		6								2.4
	2.5	8	7	6	6	7	6	6		6								2.5
	3.0	7	6	6	5	6	5	5		5								3.0
	3.5	6	5	5		5												3.5
	4.0	6																4.0

Source: Reproduced from Table E of Owen L. Davies, *The Design and Analysis of Industrial Experiments*, 2nd ed., Longman House, Essex, 1956. By permission of the author and publishers.

Note: The entries in this table show the number of observations needed in a t test of the significance of a mean in order to control the probabilities of errors of the first and second kinds at α and β respectively.

End CN Friday 11/5/21

START CN Monday 2/8/21 → able to deal w/ non normality

Robustness of t-based Procedures

The question of approximate correctness of t procedures has been studied extensively.

As was demonstrated in Handout 11, the coverage probability for confidence intervals can be quite different from the nominal (stated) level of confidence when the population distribution is heavy-tailed. A similar problem occurs in the testing situation.

When the population is very heavy-tailed, as is the case in Figure 5.21(b), the tests of hypotheses will tend to have probability of Type I errors smaller than the specified level which leads to a test having much lower power and hence greater chances of committing Type II errors.

Skewness, as depicted in Figures 5.21 (b) and (c), particularly with small sample sizes, can have an even greater effect on the probability of both Type I and Type II errors.

When we are sampling from a population distribution which has a normal distribution, the sampling distribution of a t statistic, $t = \sqrt{n}(\bar{X} - \mu)/S$, is symmetric.

However, when we are sampling from a population distribution which is highly skewed, the sampling distribution of a t statistic is skewed, not symmetric, see page 32 . Although the degree of skewness decreases as the sample size increases, there is no procedure for determining the sample size at which the sampling distribution of the t statistic becomes symmetric.

As a consequence, when testing $H_1 : \mu > \mu_o$, the level of a nominal $\alpha = .05$ test may actually have a level of .01 or less when the sample size is less than 20 and the population distribution looks like that of Figure 5.21(b), (c) or (d).

When testing $H_1 : \mu < \mu_o$, the opposite is true. That is, the level of a nominal $\alpha = .05$ test may actually have a level of .1 or larger when the sample size is less than 20 and the population distribution looks like that of Figure 5.21(b), (c) or (d).

Fig. 5.21(a) Density of the
Standard Normal Distribution

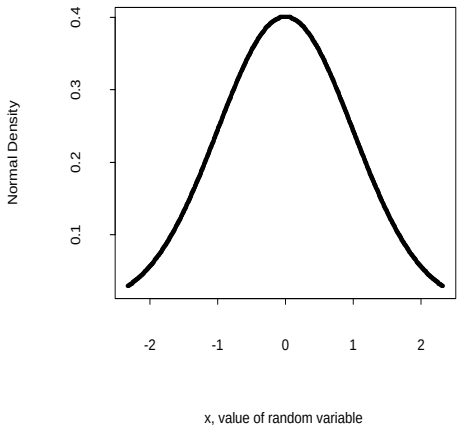


Fig. 5.21(b) Density of a
Heavy-Tailed Distribution

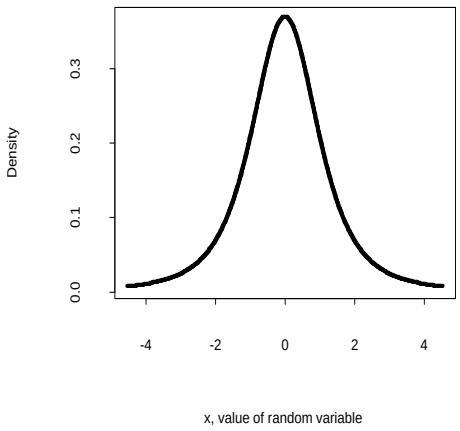


Fig. 5.21(c) Density of a
Lightly Skewed Distribution

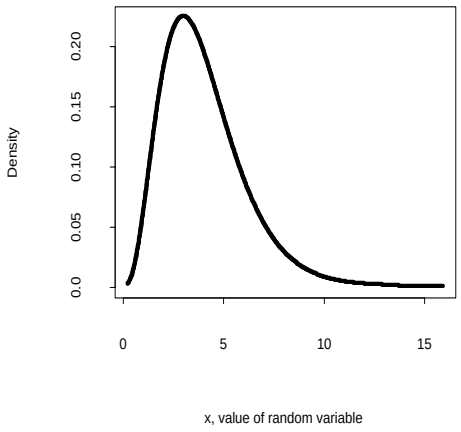
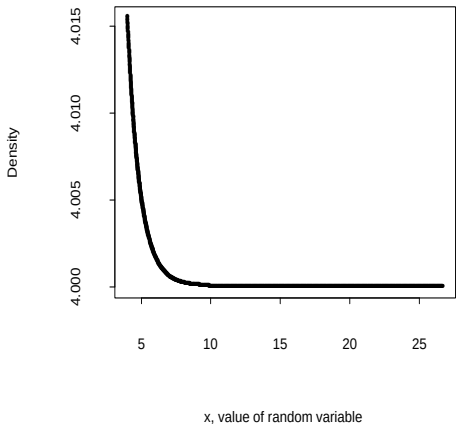


Fig. 5.21(d) Density of a
Highly Skewed Distribution



The second question, that of the *efficiency* of the t procedures, will be examined next. The power of the t test will be considerably reduced when the population distribution is symmetric but heavy-tailed thus causing an increase in the probability of Type II errors.

A simulation study of the effect of skewness and heavy-tailedness on the level and power of the t test yielded the results given in the following table.

Level and Power Values For t Test when $H_1 : \mu > \mu_o$

	n=10				n=15				n=20			
Population Distribution	0	.2	.6	.8	0	.2	.6	.8	0	.2	.6	.8
Normal	.05	.145	.543	.754	.05	.182	.714	.903	.05	.217	.827	.964
Heavy-Tailed	.035	.104	.371	.510	.049	.115	.456	.648	.045	.163	.554	.736
Light Skewness	.025	.079	.437	.672	.037	.129	.614	.864	.041	.159	.762	.935
Heavy Skewness	.007	.055	.277	.463	.006	.078	.515	.733	.011	.104	.658	.873

The values in the table are the power values for a level $\alpha = .05$ t test of H_o: $\mu \leq \mu_o$ versus H_1: $\mu > \mu_o$.

The power values are calculated for shifts of size d=$\frac{\mu_a - \mu_o}{\sigma}$=0, .2, .6, .8.

Three different sample sizes were used n=10, 15, and 20.

When d=0, this is the level of the test. We want to compare these values to .05.

The values when d>0 are compared to the corresponding values when sampling from a normal population.

We observe that when sampling from the lightly skewed distribution and the heavy-tailed distribution, the levels are somewhat less than .05 with values nearly equal to .05 when using n=20.

However, when sampling from a heavily skewed distribution, even with n=20 the level was only .011.

The power values for the heavy-tailed and heavily skewed populations are considerably less than the corresponding values when sampling from a normal distribution. Thus the test is much less likely to correctly detect that the alternative hypothesis, H_1 is true. This reduced power is present even when n=20.

When sampling from a lightly skewed population distribution, the power values are very nearly the same as the values for the normal distribution.

Since the t procedures have reduced power when sampling from skewed populations with small sample sizes, procedures have been developed which are not as affected by the skewness or extreme heavy tailedness of the population distribution.

These procedures are called **robust methods** of estimation and inference. Two robust procedures, the sign test and Wilcoxon signed rank test will be considered. They are both more efficient than the t test when the population distribution is very nonnormal in shape. Also, they maintain the selected α level of the test unlike the t test which when applied to very nonnormal data has a true α value much different from the selected α value.

When testing $H_1 : \mu < \mu_o$, the true size of the test is larger than the nominal size when the sample is from highly skewed distribution. This results in greater power but at the risk of having the probability of Type I error much larger than what is stated.

The same comments can be made with respect to confidence intervals for the mean.

When the population distribution is highly skewed, the coverage probability of a nominal $100(1-\alpha)$ confidence interval will be considerably less than $100(1-\alpha)$.

Thus, we can conclude that the size of tests for two-sided alternatives, $H_1 : \mu \neq \mu_o$, will be larger than stated.

In practice, how should we interpret these results. First, examine the data through graphs. A box plot or normal probability plot will reveal any gross skewness or extreme outliers.

If the plots do not reveal extreme skewness or many outliers, the nominal t-distribution probabilities should be reasonably correct. Thus, the level and power calculations for tests of hypotheses, and the coverage probability of confidence intervals should be reasonably accurate.

If the plots reveal severe skewness or heavy-tailedness, the test procedures and confidence intervals based on the t-distribution will be highly suspect.

In these situations, the median is a more appropriate measure of the center of the population than is the mean. We will develop test of hypotheses and confidence intervals for the median of a population. These procedures will maintain the nominal coverage probability of confidence intervals and the stated α level of tests of hypotheses when the population distribution is highly skewed or heavy-tailed.

A plot of the sampling distribution of the $t-test$ when taking samples of size $n = 10, 20, 30$ from a Gamma($\alpha = .5, \beta = 1$) distribution is given on the next page.

The plots were simulated using 10,000 replications. The estimated percentiles are

		Percentiles of t-test statistics			
n	Population Distribution	.025	.05	.95	.975
10	Normal	-2.262	-1.833	1.833	2.262
10	Gamma(.5,1)	-5.829	-4.167	1.255	1.505
20	Normal	-2.093	-1.729	1.729	2.093
20	Gamma(.5,1)	-3.997	-3.085	1.309	1.530
30	Normal	-2.045	-1.699	1.699	2.045
30	Gamma(.5,1)	-3.291	-2.575	1.337	1.581

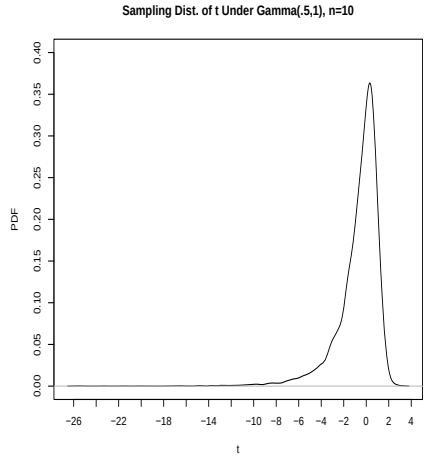
The above percentiles demonstrate why the size of the test is higher than the nominal size for two sided and one-sided lower hypotheses and is lower than the nominal size for one-sided higher hypotheses when the population distribution is highly right skewed. Suppose data is from Gamma(.5, .1) distribution.

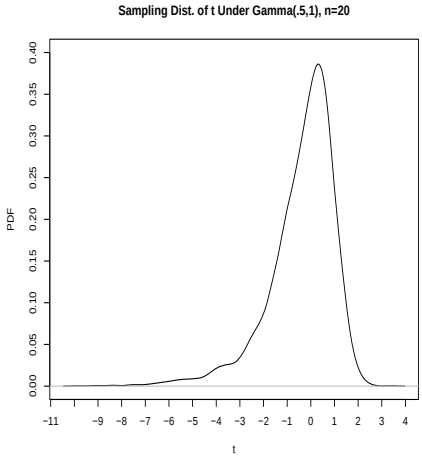
To test $H_1 : \mu > \mu_o$ with n = 10, $\alpha = .05$. The researcher uses TS = $\frac{(\bar{X}-\mu_o)}{S/\sqrt{n}}$ with decision rule Reject H_o if TS> $t_{.05,9} = 1.833$. What is the true level of the test?

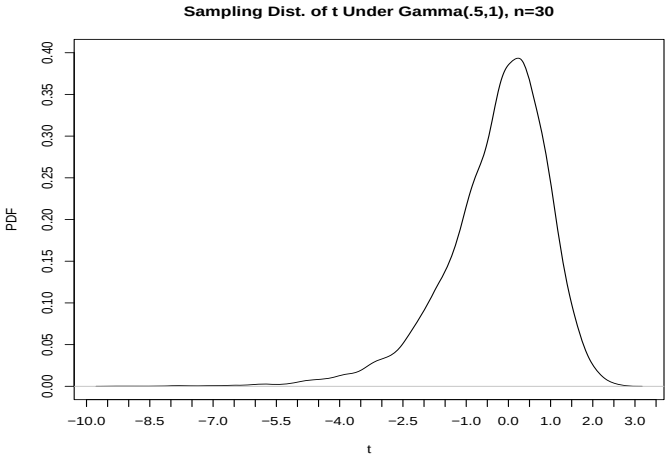
$P[TS > 1.833] < P[TS > 1.255] = .05$. Thus, true level of the test is less than .05 which produces a test having P[Type II error] greater than the value expected if data was from a normal distribution.

To test $H_1 : \mu < \mu_o$ with n = 10, $\alpha = .05$. The researcher uses TS = $\frac{(\bar{X}-\mu_o)}{S/\sqrt{n}}$ with decision rule Reject H_o if TS< $-t_{.05,9} = -1.833$. What is the true level of the test?

$P[TS < -1.833] > P[TS < -4.167] = .05$. Thus, true level of the test is greater than .05 which produces a test having P[Type I error] greater than the value expected if data was from a normal distribution.


Sampling Dist. of t Under Gamma(.5,1), n=10
PDF
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−26
−22
−18
−14
−10
−8
−6
−4
−2
0
2
4
t


Sampling Dist. of t Under Gamma(.5,1), n=20
PDF
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−11
−9
−8
−7
−6
−5
−4
−3
−2
−1
0
1
2
3
4
t


Sampling Dist. of t Under Gamma(.5,1), n=30
PDF
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−10.0
−8.5
−7.0
−5.5
−4.0
−2.5
−1.0
0.0
1.0
2.0
3.0
t

Correlated Data

What happens to the t-test if the data is not independent but is correlated:

1. Equicorrelated: $Cov(X_i, X_j) = \rho\sigma^2 \quad \text{for all} \quad i \neq j \quad \frac{-1}{n-1} < \rho < 1 \quad \Rightarrow$

   $E[\bar{X}] = \mu \quad \text{and} \quad Var(\bar{X}) = \frac{\sigma^2}{n}[1 + (n-1)\rho]$

2. 1st Order Autoregressive: $Cov(X_i, X_j) = \rho^{|i-j|}\sigma^2 \quad \text{for} \quad i \neq j \quad 1 < \rho < 1 \quad \Rightarrow$

   $E[\bar{X}] = \mu \quad \text{and} \quad Var(\bar{X}) \approx \frac{\sigma^2}{n}\left[\frac{1+\rho}{1-\rho}\right]$

In both cases, if $\rho > 0$, then $Var(\bar{X}) > \frac{\sigma^2}{n}$. Thus, $\frac{S}{\sqrt{n}}$ underestimates $SE(\bar{X})$.

This results in a C.I. $\bar{X} \pm t_{\alpha/2}\frac{S}{\sqrt{n}}$ which is too narrow and hence the coverage probability is less than $100(1-\alpha)\%$.

Also, the test statistic $\frac{\sqrt{n}(\bar{X}-\mu_o)}{S}$ is larger than what it should be if the correct value for the estimated standard error was used in the denominator.

Thus, the probability of Type I error is inflated above α, resulting in an inflated proportion of Type I errors.

Suppose we are testing $H_1 : \mu > \mu_o$ and the data is positively correlated. Thus,

$S/\sqrt{n} < \widehat{SE}(\bar{X}) \Rightarrow \frac{\bar{X}-\mu_o}{S/\sqrt{n}} > \frac{\bar{X}-\mu_o}{\widehat{SE}(\bar{X})} \Rightarrow$

P[Type I error] $= P[\frac{\bar{X}-\mu_o}{S/\sqrt{n}} > t_{\alpha,n-1}] > P[\frac{\bar{X}-\mu_o}{\widehat{SE}(\bar{X})} > t_{\alpha,n-1}] = \alpha$

Thus, the probability of a Type I error is larger than the specified value when postive correlation is present and the test statistics is not adjusted for the correlation. However, the power of the test is also inflated but this is paid for by the inflated Type I error rate.

We need to detect when correlation is present and adjust C.I.'s and tests of hypotheses for the correlation.

If possible we need to determine the type of correlation present and then estimate the standard error of $\bar{X}$ to take into account the correlation in the data.

The critical value of the test statistic would need to be adjusted also.

Time series STAT 626, multivariate analysis STAT 636, and spatial statistics STAT 647 deal with these types of issues.

Case 3: NonNormally Distributed Population

When the sample size n is large and population distribution is not $N(\mu, \sigma^2)$, the sampling distribution of the test statistic $\sqrt{n}(\bar{X} - \mu_o)/S$ is approximately represented by a t-distribution. We can just apply the Central Limit Theorem. How large n needs to be to obtain an appropriate approximation depends on the shape of the population/process distribution. The more skewed or heavy-tailed the population distribution is relative to a normal distribution, the greater the required sample size.

When the sample size n is small and population distribution is not $N(\mu, \sigma^2)$, the sampling distribution of the test statistic $\sqrt{n}(\bar{X} - \mu_o)/S$ is not adequately represented by a t-distribution. This was demonstrated on pages 28-32.

Thus, for small sample sizes we need to select alternative methods.

Suppose we wanted to test $H_o : \mu \leq \mu_o$ versus $H_1 : \mu > \mu_o$.

We obtain a random sample $Y_1, \ldots, Y_n$ which are iid with mean μ but the cdf is not a normal cdf.

We could use the Box-Cox transformation and potentially obtain a transformation $X_i = g(Y_i)$ and hence the data set $X_1, \ldots, X_n$ would be approximately normally distributed.

We could then test $H_o : \mu \leq \mu_o$ versus $H_1 : \mu > \mu_o$ but would need to restate the hypotheses in terms of the mean of the X_i and the test statistic would become:

$$t = \frac{\sqrt{n}(\bar{X} - \mu_o^*)}{S_X}$$

What value should we use for μ_o^*, the mean of X when the mean of Y is μ_o?

Would $\mu_o^* = g(\mu_o)$ be an appropriate value?

In some situations yes, depending of the appropriateness of using a 1-term Taylor series expansion of $g(\cdot)$ about μ_o.

In many cases, this leads to test statistics that have size greatly different from α due to the misidentification of the mean of X_i.

Therefore, transformations in this situation are not generally recommended.

For small to moderate sample sizes, two alternatives exist:

Sign Test or **Wilcoxon Signed-Rank Test**:

Sign Test - Hypotheses about the Population Median

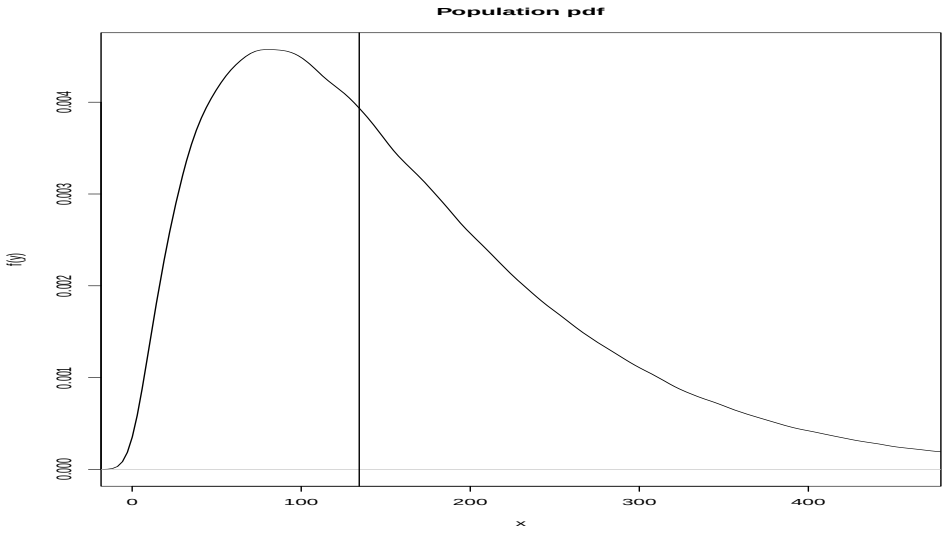
Let $Y_1, \ldots, Y_n$ be iid with continuous cdf F and median $\tilde{\mu}$.

Note: If population distribution is symmetric then $\mu = \tilde{\mu}$. Therefore, hypotheses about the median, $\tilde{\mu}$, are the same as tests about the mean, μ. When the population distribution is highly skewed, we will use bootstrap methods to conduct tests of hypotheses about the population mean.

To test hypotheses about $\tilde{\mu}$ relative to $\tilde{\mu}_o$ we can use the following test statistic:

- Let $X_i = Y_i - \tilde{\mu}_o$ and
- $S_+ = \sum_{i=1}^{n} I(X_i > 0)$, number of positive X_i or number of Y_i's greater than $\tilde{\mu}_o$.
- S_+ has a $Bin(n, p)$ distribution with
- $p = P[X_i > 0] = P[Y_i > \tilde{\mu}_o]$.

  Thus, we just convert a test about $\tilde{\mu}$ to a test about p:
- $\tilde{\mu} > \tilde{\mu}_o \Rightarrow \quad p = P[X_i > 0] = P[Y_i > \tilde{\mu}_o] > \frac{1}{2}$
- $\tilde{\mu} = \tilde{\mu}_o \Rightarrow \quad p = P[X_i > 0] = P[Y_i > \tilde{\mu}_o] = \frac{1}{2}$
- $\tilde{\mu} < \tilde{\mu}_o \Rightarrow \quad p = P[X_i > 0] = P[Y_i > \tilde{\mu}_o] < \frac{1}{2}$



- Note: The sampling distribution of the sign test is binomial only if the cdf of Y_i is continuous.

  Thus, $P[Y_i = \tilde{\mu}_o] = 0$. However, in actual data, due to rounding off of measured values of Y_i we will occasionally have $Y_i = \tilde{\mu}_o$.

  In these cases, reduce the sample size to $n^* = n - k$ where k is the number of $Y_i = \tilde{\mu}_o$. That is, n^* is the number of observations not equal to $\tilde{\mu}_o$.

The following three sets of hypotheses will now be examined.

H1. For $H_0 : \tilde{\mu} \leq \tilde{\mu}_o$ versus $H_1 : \tilde{\mu} > \tilde{\mu}_o \Rightarrow \quad H_1 : p > \frac{1}{2}$,

- Reject H_o if $S_+ \geq B_{n,\alpha} = qbinom(1-\alpha, n, .5)$,

  the upper α percentile of the $B(n, .5)$ distribution.

- $p - value = P[B(n, .5) \geq S_+] = 1 - G(S_+ - 1) = 1 - pbinom(S_+ - 1, n, .5)$,
  where G is the cdf of a $B(n, .5)$

- The power function is given by

  $\gamma(\tilde{\mu}_1) = P[S_+ \geq B_{n,\alpha} | \tilde{\mu} = \tilde{\mu}_1] = 1 - G_{p_1}(B_{n,\alpha} - 1) = 1 - pbinom(qbinom(1-\alpha, n, .5) - 1, n, p_1)$,

  where G_{p_1} is the cdf of a $B(n, p_1)$ and

$$p_1 = P[Y_i > \tilde{\mu}_o | \tilde{\mu} = \tilde{\mu}_1] = \int_{\tilde{\mu}_o}^{\infty} f_Y(y)dy$$

  where $f_Y(y)dy$ is the pdf of Y with median $\tilde{\mu}_1$. Thus, the population pdf must be known in order to calculate the power of the Sign Test.

H2. For $H_0 : \tilde{\mu} \geq \tilde{\mu}_o$ versus $H_1 : \tilde{\mu} < \tilde{\mu}_o \Rightarrow \quad H_1 : p < \frac{1}{2}$,

- Reject H_o if $S_+ \leq B_{n,1-\alpha} = qbinom(\alpha, n, .5)$,

  the lower α percentile of the $B(n, .5)$ distribution.

- $p - value = P[B(n, .5) \leq S_+] = G(S_+) = pbinom(S_+, n, .5)$,

  where G is the cdf of a $B(n, .5)$

- The power function is given by

  $\gamma(\tilde{\mu}_1) = P[S_+ \leq B_{n,1-\alpha} | \tilde{\mu} = \tilde{\mu}_1] = G_{p_1}(B_{n,1-\alpha}) = pbinom(qbinom(\alpha, n, .5), n, p_1)$,

  where G_{p_1} is the cdf of a $B(n, p_1)$ and

$$p_1 = P[Y_i > \tilde{\mu}_o | \tilde{\mu} = \tilde{\mu}_1] = \int_{\tilde{\mu}_o}^{\infty} f_Y(y)dy$$

  where $f_Y(y)dy$ is the pdf of Y with median $\tilde{\mu}_1$.

H3. For $H_0 : \tilde{\mu} = \tilde{\mu}_o$ versus $H_1 : \tilde{\mu} \neq \tilde{\mu}_o \Rightarrow \quad H_1 : p \neq \frac{1}{2}$,

Let $S_{max} = max(S_+, n - S_+)$

- Reject H_o if $S_{max} \geq B_{n,\alpha/2} = qbinom(1 - \alpha/2, n, .5)$,

  where $B_{n,\alpha/2}$ is the upper $\alpha/2$ percentile of the $B(n, .5)$ distribution.

- $p - value = 2P[B(n, .5) \geq S_{max}] = 2(1 - G(S_{max} - 1)) = 2(1 - pbinom(S_{max} - 1, n, .5)$,

  where G is the cdf of a $B(n, .5)$

- The power function is given by

$$\begin{aligned}
\gamma(\tilde{\mu}_1) &= P[S_{max} \geq B_{n,\alpha/2} | \tilde{\mu} = \tilde{\mu}_1] \\
&= P[S_+ \geq B_{n,\alpha/2} | \tilde{\mu} = \tilde{\mu}_1] + P[S_+ \leq n - B_{n,\alpha/2} | \tilde{\mu} = \tilde{\mu}_1] \\
&= 1 - G_{p_1}(B_{n,\alpha/2} - 1) + G_{p_1}(n - B_{n,\alpha/2}) \\
&= 1 - pbinom(qbinom(1 - \alpha/2), n, .5) - 1, n, p_1) + pbinom(n - qbinom(1 - \alpha/2, n, .5), n, p_1)
\end{aligned}$$

where G_{p_1} is the cdf of a $B(n, p_1)$ and

$p_1 = P[Y_i > \tilde{\mu}_o | \tilde{\mu} = \tilde{\mu}_1] = \int_{\tilde{\mu}_o}^{\infty} f_Y(y)dy$

where $f_Y(y)dy$ is the pdf of Y with median $\tilde{\mu}_1$.

Comments:

- Based on the form of the Rejection Region (RR), p-value, and power function, we can conclude that

  1. The Sign test is **distribution-free** when $\tilde{\mu} = \tilde{\mu}_o$ because
     the RR and p-value do not depend on the form of the population pdf
  2. When $\tilde{\mu} \neq \tilde{\mu}_o$, the Sign test is **not distribution-free**, because the power function depends on the form of the population pdf.

- One problem with the sign test is that it is based on the binomial distribution and hence the exact level of the test cannot be guaranteed due to the discreteness of the binomial distribution. The following table shows the exact level of the sign test for various values of n and nominal α

n	α	true level	n	α	true level	n	α	true level
10	0.05	0.055	10	0.01	0.011	10	0.001	0.0107
25	0.05	0.054	25	0.01	0.022	25	0.001	0.00204
50	0.05	0.059	50	0.01	0.016	50	0.001	0.00130
75	0.05	0.053	75	0.01	0.010	75	0.001	0.00122
100	0.05	0.067	100	0.01	0.010	100	0.001	0.00176
500	0.05	0.059	500	0.01	0.011	500	0.001	0.00100
1000	0.05	0.053	1000	0.01	0.010	1000	0.001	0.00107

$I_i = \begin{cases} 1 & \text{if } Y_i \geq \tilde{\mu}_0 \\ 0 & \text{if } Y_i < \tilde{\mu}_0 \end{cases} \Rightarrow$ loss of information b/c we don't know how much greater than. Re location fixes issue.

STOP Monday 9/8/21

START Wednesday 11/10/2021

Wilcoxon Signed-Rank Test

→ added condition corrected to above.

Let $Y_1, \ldots, Y_n$ be iid with continuous symmetric pdf f and median $\tilde{\mu}$.

Because the population distribution is symmetric then, tests about the median are the same as tests about the mean.

To test hypotheses about $\tilde{\mu}$ relative to $\tilde{\mu}_o$ we can use the following procedure:

1. Let $X_i = Y_i - \tilde{\mu}_o$
2. Discard any $X_i = 0$ and let n^* be the number of nonzero X_i
3. Rank the magnitude of the X_i's: $|X_i| = |Y_i - \tilde{\mu}_o|$ from smallest to largest
4. In the case of ties, assign the average of the ranks to each member of the tied group
5. Let $W_+ =$ sum of the ranks associated with $X_i > 0$ and $W_- =$ sum of the ranks associated with $X_i < 0$ (If there is no ties, then $W_+ + W_- = \sum_{i=1}^{n} i = n(n+1)/2$)
6. Use Table in eCampus under Lecture Notes or R, to obtain the critical values $W_{n,\alpha}$ for each of the following cases:

A. $H_o : \tilde{\mu} \leq \tilde{\mu}_o$ versus $H_1 : \tilde{\mu} > \tilde{\mu}_o$

Reject H_o if $W_+ \geq W_{n,\alpha} = qsignrank(\alpha, n, FALSE)$

True level of the test is $psignrank(qsignrank(\alpha, n, FALSE) - 1, n, FALSE)$

$p - value = P[W \geq W_+] = P[W > W_+ - 1] = psignrank(W_+ - 1, n, FALSE)$, or use values given in Table A10.

B. $H_o : \tilde{\mu} \geq \tilde{\mu}_o$ versus $H_1 : \tilde{\mu} < \tilde{\mu}_o$

Reject H_o if $W_+ \leq W_{n,1-\alpha} = qsignrank(\alpha, n, TRUE)$

True level of the test is $psignrank(qsignrank(\alpha, n, TRUE), n, TRUE)$

$p - value = P[W \leq W_+] = psignrank(W_+, n, TRUE)$, or use values given in Table A10.

C. $H_o : \tilde{\mu} = \tilde{\mu}_o$ versus $H_1 : \tilde{\mu} \neq \tilde{\mu}_o$

Let $W_{max} = max(W_+, W_-)$

Reject H_o if $W_{max} \geq W_{n,\alpha/2} = qsignrank(\alpha/2, n, FALSE)$

True level of the test is $2psignrank(qsignrank(\alpha/2, n, FALSE) - 1, n, FALSE)$

$p - value = 2P[W \geq W_{max}] = 2P[W > W_{max} - 1] = 2psignrank(W_{max} - 1, n, FALSE)$, or

use values given in Table A10.

Note: The R-function **psignrank(x,n,FALSE)** $= P[W > x]$ but

psignrank(x,n,TRUE) $= P[W \leq x]$

qsignrank(α**,n,TRUE)** is the lower α percentile of the Wilcoxon statistics whereas

qsignrank(α**,n,FALSE)** is the upper α percentile

Wilcoxon Signed Rank Test in R

The R-function given below can be used to conduct the Wilcoxon signed rank (and Wilcoxon Rank Sum Test):

```
wilcox.test(x, y = NULL, alternative = c("two.sided", "less", "greater"),
                  mu = 0, paired = FALSE, exact = NULL, correct = TRUE,
                 conf.int = FALSE, conf.level = 0.95, ...)
```

1. $x = (x_1, x_2, \ldots, x_n)$;
2. $y = (\tilde{\mu}_o, \tilde{\mu}_o, \ldots, \tilde{\mu}_o)$;
3. set **paired=T**;
4. set **alternative** depending on the alternative hypotheses.

- Note: if **paired=T** then we have Wilcoxon Signed Rank Test and if **paired=F** then we have Wilcoxon Rank Sum Test

Asymptotic Version of Wilcoxon Signed Rank Test

For large n, we can use an asymptotic result that states:

- When $\tilde{\mu} = \tilde{\mu}_o$, the distribution of

$$\frac{W_+ - \mu_W}{\sigma_W} \text{ converges in distribution to } N(0,1),$$

where $\mu_W = \frac{n(n+1)}{4}$ and $\sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{24}}$.

- Therefore, we can use the $N(0,1)$ distribution to set the critical values and to compute $p - values$ when n is large.

$W_{n,\alpha} \approx \mu_W + Z_\alpha \sigma_W$

For testing $H_o : \tilde{\mu} \leq \tilde{\mu}_o$ versus $H_1 : \tilde{\mu} > \tilde{\mu}_o$

$p - value = P[W \geq W_+] \approx P\left[Z \geq \frac{W_+ - \mu_W}{\sigma_W}\right] = 1 - pnorm\left(\frac{W_+ - \mu_W}{\sigma_W}\right)$

- The power function of W_+ depends on the population cdf and a rather complex probability computation depending on the possible rankings under the alternatives. The book *Nonparametric Statistical Methods*, by M. Hollander and D. Wolfe, have an approximation to the power function. This then allows them to compute approximations to obtaining appropriate sample sizes for given power and size specifications.

Comparison of t-Test, Sign Test, and Wilcoxon Signed Rank Test

1. The t-Test is designed for $N(\mu, \sigma^2)$ data and tests hypotheses about the mean μ and equivalently the median $\tilde{\mu}$.
2. The Sign Test is designed for any continuous distribution and tests hypotheses about the median,$\tilde{\mu}$.
3. The Wilcoxon Signed Rank Test is designed for **symmetric** continuous distributions and tests hypotheses about the median $\tilde{\mu}$ which is equal to the mean μ when it exists.

How crucial are the conditions placed on these test statistics? The results from a large simulation study discussed in the book *Introduction to the Theory of Nonparametric Statistics*, by R. Randles and D. Wolfe provides us with the following results:

The Monte Carlo study involved 5000 replications from five distributions.

The values listed in the table are for a test of size $\alpha = .05$.

The power was simulated for shifts in the median of size

$\theta = 0\sigma, \quad .2\sigma, \quad .4\sigma, \quad .6\sigma, \quad and \quad .8\sigma,$

where σ denotes the standard deviation, of the distribution.

When the population distribution is a Cauchy centered at 0,

σ denotes the value such that the probability between $-\sigma$ and σ is the same as the probability between -1 and 1 for a standard normal distribution.

		UNIFORM					NORMAL					DOUBLE EXP					CAUCHY				
		θ/σ					θ/σ					θ/σ					θ/σ				
n	TEST	0	.2	.4	.6	.8	0	.2	.4	.6	.8	0	.2	.4	.6	.8	0	.2	.4	.6	.8
10	T^+	.051	.136	.294	.512	.746	.049	.146	.330	.543	.758	.047	.172	.374	.602	.781	.028	.095	.197	.309	.414
	W^+	.049	.136	.277	.474	.681	.050	.144	.315	.527	.741	.048	.190	.412	.633	.804	.049	.166	.332	.493	.623
	B	.049	.101	.188	.303	.453	.054	.132	.263	.440	.633	.048	.197	.407	.617	.782	.047	.183	.390	.579	.720
15	T^+	.051	.169	.408	.703	.914	.048	.181	.424	.716	.906	.049	.202	.473	.739	.898	.025	.094	.210	.321	.418
	W^+	.051	.163	.383	.642	.852	.047	.178	.418	.693	.893	.050	.226	.532	.786	.926	.050	.196	.423	.622	.750
	B	.053	.124	.230	.390	.590	.048	.149	.331	.564	.777	.051	.239	.528	.775	.907	.052	.228	.498	.733	.861
20	T^+	.049	.214	.522	.829	.971	.048	.225	.546	.833	.967	.044	.238	.571	.835	.955	.026	.099	.214	.329	.433
	W^+	.050	.205	.479	.768	.935	.049	.218	.531	.813	.962	.049	.284	.652	.885	.975	.049	.234	.514	.730	.849
	B	.055	.133	.278	.487	.703	.056	.186	.417	.677	.873	.052	.297	.644	.869	.962	.055	.274	.608	.831	.930

Tests: T^+ = Student t-test; W^+ = Wilcoxon signed rank test; B = Sign test

The entries in the table are the estimated probability of rejecting H_o, $\hat{P} = \hat{\gamma}(\theta/\sigma)$

The standard error of the estimated P is $SE(\hat{P}) = \sqrt{\hat{P}(1-\hat{P})/5000}$

$SE(\hat{P}) = .0031$ for P near .05 or .95; $SE(\hat{P}) = .0057$ for P near .2 or .8;

$SE(\hat{P}) = .0067$ for P near .35 or .65; $SE(\hat{P}) = .0071$ for P near .5

From the results of the simulation, we reach the following conclusions:

1. For a population having a normal distribution, the t-test is only slightly better than the Wilcoxon signed rank test.

2. For a population having a normal distribution, the sign test is not competitive.

3. When the population distribution is double exponential, the sign test is best for small shifts, but the Wilcoxon signed rank test is better for large shifts in the median. The t-test is competitive but third.

4. With very heavy tailed population distributions, for example, the Cauchy distribution, the t-test becomes very conservative (the true size is much lower than .05), and its power drops off considerably from its value under a normal distribution. The sign test has greatest power with the signed rank test having somewhat smaller power values.

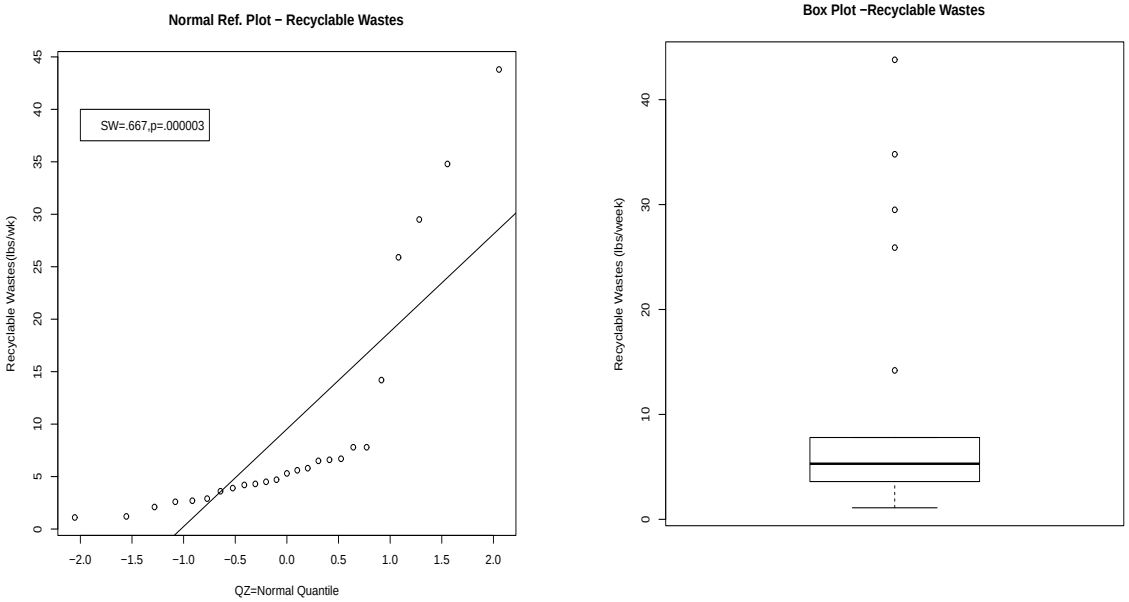
Example of the Application of t-Test, Wilcoxon Signed Rank Test, and Sign Test

The sanitation department of a large city wants to investigate ways to reduce the amount of recyclable materials that are placed in the city's landfill. By separating the recyclable material from the remaining garbage, the city could prolong the life of the landfill site. More importantly, the number of trees needed to be harvested for paper products and aluminum needed for cans could be greatly reduced. From an analysis of recycling records from other cities, it is determined that if the average weekly amount of recyclable material is more than five pounds per household a commercial recycling firm could make a profit collecting the material. To determine the feasibility of the recycling plan, 25 households were randomly selected for inclusion in the study. The weekly weight of recyclable material (pounds/week) for each household is given here:

14.2	5.3	2.9	4.2	1.2	4.3	1.1	2.6	6.6	7.8	25.9	43.8	2.7
5.6	7.8	3.9	4.7	6.5	29.5	2.1	34.8	3.6	5.8	4.5	6.7	

The Sanitation Department wanted to determine whether the average household recyclable wastes was greater than 5 pounds per week. Test this research hypothesis at the $\alpha = .05$ level.

Because $n = 25$ is a relatively small sample size, we need to determine the most appropriate measure of "average". A plot of the data is given next.



From the plots, the data appears to be very right skewed and definitely not from a normal distribution (Shapiro-Wilk test have p-value=.000003). Therefore, the mean would not be an appropriate measure of the average amount of recyclable wastes per household.

The median, $\tilde{\mu}$, would be a better measure of the average than would be the mean because of the heavy skewness in the data.

The set of hypotheses for testing the Sanitation Department's research hypothesis is

$$H_o : \tilde{\mu} \leq 5 \quad \text{versus} \quad H_1 : \tilde{\mu} > 5$$

The order statistics for the data set, $(Y_{(1)}, \ldots, Y_{(25)})$, are given here:

1.1	1.2	2.1	2.6	2.7	2.9	3.6	3.9	4.2	4.3	4.5	4.7	5.3
5.6	5.8	6.5	6.6	6.7	7.8	7.8	14.2	25.9	29.5	34.8	43.8	

The sample median value is $\hat{\tilde{\mu}} = Y_{(13)} = 5.3$. Therefore, there is some evidence in the data to support $H_1 : \tilde{\mu} > 5$. We will now contact a statistical test.

Because the population distribution appears to be not normally distributed the appropriate test statistics would be the sign test or the Wilcoxon Signed Rank Test. We will conduct the Sign Test first.

Sign Test

Let S_+ be the number of data values Y_is greater than $\tilde{\mu}_o = 5$. our critical value.

Determine S_{cr} such that $P[B \geq S_{cr}] \leq .05$ and $P[B \geq S_{cr} - 1] > .05$

With B distributed $B(25, .5)$, $Pr[B \geq 18] = 1 - pbinom(17, 25, .5) = 0.023$, and

$Pr[B \geq 17] = 1 - pbinom(16, 25, .5) = .054$.

Therefore, Reject H_o if $S_+ \geq 18$. in order to have the size of the test less than or equal to .05.

The actual α is thus 0.023, not 0.05

Using R, $qbinom(.95, 25, .5) = 17$ which would yield a level .054 test not .05.

From the data there are 13 values for Y_i which are greater than 5, thus $S_+ = 13 < 18 = S_{cr} \Rightarrow$ Fail to reject H_o

Let B have a Binomial distribution with $p = .5$ and $n = 25$

$$\begin{aligned} p - value \;=\; Pr[B \geq S_+] \;=\; Pr[B \geq 13] \quad &= \quad 1 - Pr[B \leq 12] \\ &= \quad 1 - pbinom(12, 25, .5) \\ &= \quad 0.5 > 0.05 = \alpha \end{aligned}$$

Thus, fail to reject H_o.

Conclusion, there is insufficient evidence, p-value=.5, that the median level of household recyclable wastes is greater than 5 pounds.

A 95.7% C.I. on the median recyclable wastes, $\tilde{\mu}$ is obtained from Table VII.3 in Handout 11.

Taking k=8, the 95.7% C.I. for the population median is

$$\left(Y_{(k)},\ Y_{(n-k+1)}\right) \;=\; \left(Y_{(8)},\ Y_{(18)}\right) \;=\; (3.9,\ 6.7)$$

Sign Test Using R

R has a function **binom.test(x,n,p=.5,alternative="greater",conf.level=.95)**

which will test both a two-sided research hypothesis:

$H_1 : \tilde{\mu} \neq \tilde{\mu}_o$, **alternative="two.sided"** and

a one-sided research hypothesis:

$H_1 : \tilde{\mu} > \tilde{\mu}_o$, **alternative="greater"**

In our example, we had $\tilde{\mu}_o = 5$ with 13 of the n = 25 data values greater than 5. Thus, we can use the following R code to test the hypotheses:

$H_0 : \tilde{\mu} \leq 5$ versus $H_1 : \tilde{\mu} > 5$ or equivalently

$H_o : p \leq \frac{1}{2}$ versus $H_1 : p > \frac{1}{2}$

where p is the proportion of the population greater than $\tilde{\mu}_0 = 5$

```
 binom.test(13,25,p=.5,"greater")

        Exact binomial test

data:  13 and 25

number of successes = 13, number of trials = 25,

p-value = 0.5

alternative hypothesis: true probability of success is greater than 0.5

95 percent confidence interval on p: (Clopper-Pearson)

 0.3413887 1.0000000

sample estimates:

probability of success = 13/25 = 0.52
```

Alternative Test Statistic: Wilcoxon Signed-Rank Test

For testing,

$$H_0 : \tilde{\mu} \leq 5 \quad \text{versus} \quad H_1 : \tilde{\mu} > 5$$

Let W_+ be sum of the ranks associated with the positive values of $X_i = Y_i - 5$:

Then, Reject H_o if $W_+ \geq W_{25,.05} \approx \frac{(25)(26)}{4} + z_{.05}\sqrt{\frac{(25)(26)(51)}{24}} = 223.6$

Asymptotic approximation.

Let $X_i = Y_i - 5$ and rank the values of $|X_i|$:

X_i	-3.9	-3.8	-2.9	-2.4	-2.3	-2.1	-1.4	-1.1	-0.8	-0.7	-0.5	-0.3	0.3
Rank	20	19	18	15	14	13	9	8	6.5	5	3	1.5	1.5

X_i	0.6	0.8	1.5	1.6	1.7	2.8	2.8	9.2	20.9	24.5	29.8	38.8
Rank	4	6.5	10	11	12	16.5	16.5	21	22	23	24	25

Summing the ranks associated with positive X_i yields:

$W_+ = 1.5 + 4 + 6.5 + 10 + 11.5 + 11.5 + 16.5 + 16.5 + 21 + 22 + 23 + 24 + 25 = 193 < 223.6 \quad \Rightarrow$

Fail to RejectH_o

$$p - value = Pr[W_+ \geq 193] \approx 1 - \Phi\left(\frac{(193) - (25)(26)/4}{\sqrt{\frac{(25)(26)(51)}{24}}}\right) = 0.2058 > 0.05 = \alpha$$

Thus, the Wilcoxon Signed Rank Test has a somewhat smaller p-value that the Sign Test, but for any sensible value of α, the two tests would reach the same conclusion.

Using the critical values from R, we would have

Reject H_o if $W_+ \geq qsignrank(.05, 25, F) = 224$ with

p-value $= P[W_+ \geq 193] = psignrank(193 - 1, 25, FALSE) = .2131$

Analysis Using R

Using the R-function **wilcox.test(y,c,alternative="g",paired=T,conf.int=T)**

with

$y = c(14.2, 5.3, 2.9, 4.2, 1.2, 4.3, 1.1, 2.6, 6.6, 7.8, 25.9,$
$43.8, 2.7, 5.6, 7.8, 3.9, 4.7, 6.5, 29.5, 2.1, 34.8, 3.6, 5.8, 4.5, 6.7)$

and c a vector of length n=25, all values=5, $c = rep(5, 25)$,

we obtain the following:

```
        Wilcoxon signed rank test with continuity correction

data:  y and c
V = 193, p-value = 0.2097
```

t-Test

What would happen if we would have ignored that the data was very non-normal and used the t-test:

$$t = \frac{\bar{Y} - 5}{S/\sqrt{n}} = \frac{9.528 - 5}{11.34/\sqrt{25}} = 1.995$$

with p-value=$P[t_{24} > 1.995] = 0.029$

If we would have blindly applied the t-Test ignoring the non-normality of the data, our conclusion would be to reject H_o because the p-value from the t-test is $p - value = 0.029 < .05 = \alpha$. This obviously contradicts the conclusions of both the Sign Test and Wilcoxon Signed Rank Test. Thus, illustrating once again that assumptions need to be checked in order to avoid possibly incorrect conclusions.

The R-function **t.test(y,alternative="g",mu=5,conf.level=.95)** yields the following:

```
        One Sample t-test

data:  y
t = 1.9967, df = 24, p-value = 0.02866
alternative hypothesis: true mean is greater than 5
95 percent confidence interval:  5.648175      Inf
sample estimates: mean of x =     9.528
```

Tests about Normal Population Standard Deviation- σ

Let $Y_1, \ldots, Y_n$ be iid from a population/process having a $N(\mu, \sigma^2)$ distribution. The test statistic for testing hypotheses comparing σ to a specified value σ_o is

$$TS = \frac{(n-1)S^2}{\sigma_o^2},$$

which has a chi-square distribution with $df = n-1$ when $\sigma = \sigma_o$ and the population distribution is $N(\mu, \sigma^2)$. The assumption of normality is crucial with this test statistic. We will consider the three sets of hypotheses:

H1. $H_o : \sigma \leq \sigma_o$ versus $H_1 : \sigma > \sigma_o$

Reject H_o if $TS \geq \chi^2_{n-1,\alpha} = qchisq(1-\alpha, n-1)$, where $\chi^2_{n-1,\alpha}$ is the upper α percentile from a chi-square distribution with $df = n-1$

$p - value = P[\chi^2_{n-1} \geq TS] = 1 - G(TS) = 1 - pchisq(TS, n-1)$, where $G(\cdot)$ is the cdf of a chi-square distribution with $df = n-1$.

Power Function: Power at $\sigma = \sigma_1$ is given by

$$\begin{aligned}
\gamma(\sigma_1) &= P_{\sigma_1}\left[\frac{(n-1)S^2}{\sigma_o^2} \geq \chi^2_{n-1,\alpha}\right] \\
&= P_{\sigma_1}\left[\frac{(n-1)S^2}{\sigma_o^2}\left(\frac{\sigma_o^2}{\sigma_1^2}\right) \geq \frac{\sigma_o^2}{\sigma_1^2}\chi^2_{n-1,\alpha}\right] \\
&= P_{\sigma_1}\left[\frac{(n-1)S^2}{\sigma_1^2} \geq \frac{\sigma_o^2}{\sigma_1^2}\chi^2_{n-1,\alpha}\right] \\
&= 1 - G\left(\frac{\sigma_o^2}{\sigma_1^2}\chi^2_{n-1,\alpha}\right) \\
\gamma(\sigma_1) &= 1 - pchisq\left(\frac{\sigma_o^2}{\sigma_1^2}qchisq(1-\alpha, n-1), n-1\right)
\end{aligned}$$

where $G(\cdot)$ is the cdf of a chi-square distribution with $df = n-1$.

In R, p-value = 1- **G(y) = 1-pchisq(y,df) and** $\chi^2_{n-1,\alpha}$**=qchisq(**$1-\alpha$**,n-1)**

H2. $H_o : \sigma \geq \sigma_o$ versus $H_1 : \sigma < \sigma_o$

Reject H_o if $TS \leq \chi^2_{n-1,1-\alpha} = qchisq(\alpha, n-1)$, where $\chi^2_{n-1,1-\alpha}$ is the lower α percentile from a chi-square distribution with $df = n-1$

$p - value = P[\chi^2_{n-1} \leq TS] = G(TS) = pchisq(TS, n-1)$, where $G(\cdot)$ is the cdf of a chi-square distribution with $df = n-1$.

Power Function:

$$\begin{aligned}
\gamma(\sigma_1) &= P_{\sigma_1}\left[\frac{(n-1)S^2}{\sigma_o^2} \leq \chi^2_{n-1,1-\alpha}\right] \\
&= G\left(\frac{\sigma_o^2}{\sigma_1^2}\chi^2_{n-1,1-\alpha}\right) \\
&= pchisq(\left(\frac{\sigma_o^2}{\sigma_1^2}qchisq(\alpha, n-1), n-1\right)
\end{aligned}$$

where $G(\cdot)$ is the cdf of a chi-square distribution with $df = n-1$.

H3. $H_o : \sigma = \sigma_o$ versus $H_1 : \sigma \neq \sigma_o$

Reject H_o if $TS \leq \chi^2_{n-1,1-\alpha/2}$ or $TS \geq \chi^2_{n-1,\alpha/2}$

$p - value = 2min[G(TS), 1 - G(TS)] = 2min[pchisq(TS, n-1), 1 - pchisq(TS, n-1)]$, where $G(\cdot)$ is the cdf of a chi-square distribution with $df = n-1$.

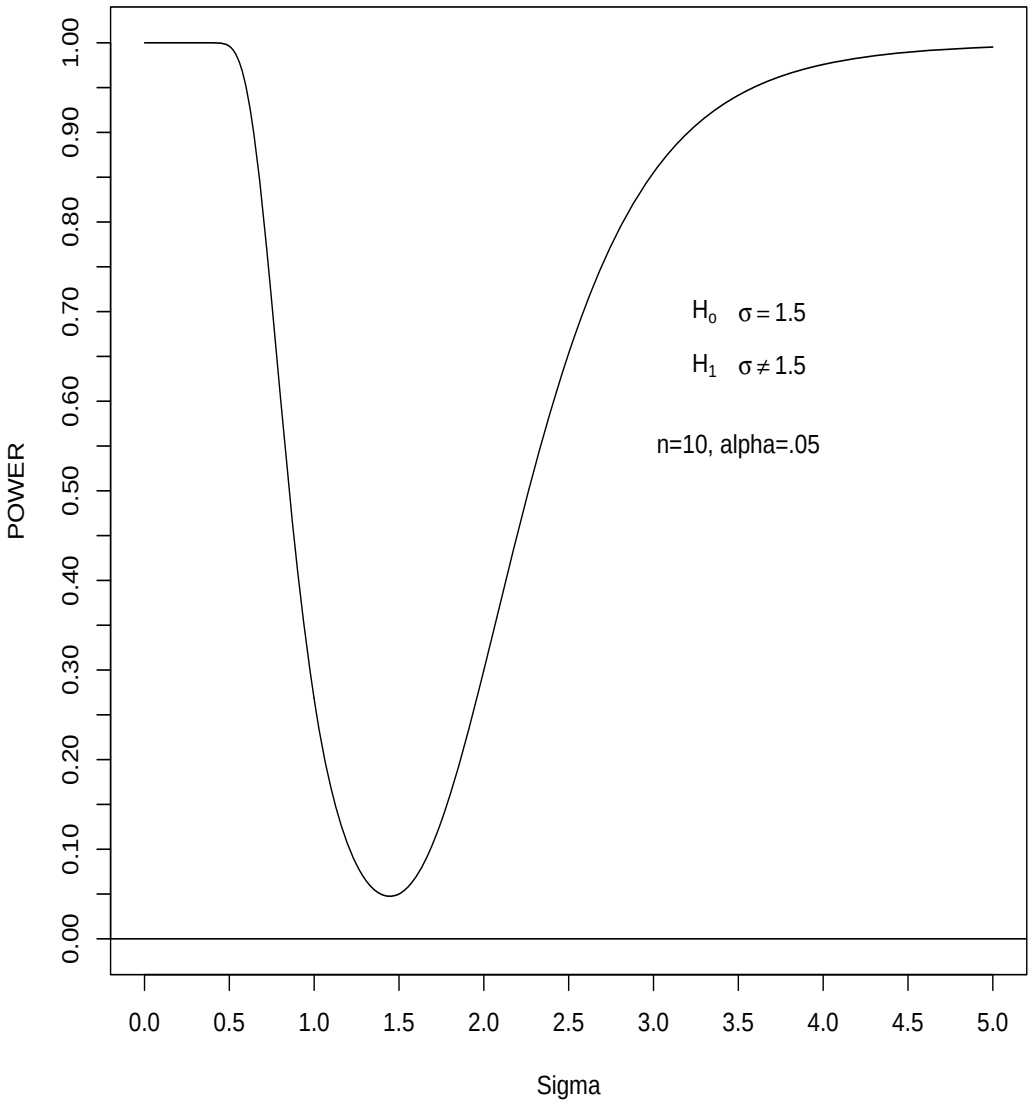
Power Function:

$$\begin{aligned}
\gamma(\sigma_1) &= P_{\sigma_1}\left[\frac{(n-1)S^2}{\sigma_o^2} \leq \chi^2_{n-1,1-\alpha/2}\right] + P_{\sigma_1}\left[\frac{(n-1)S^2}{\sigma_o^2} \geq \chi^2_{n-1,\alpha/2}\right] \\
&= G\left(\frac{\sigma_o^2}{\sigma_1^2}\chi^2_{n-1,1-\alpha/2}\right) + 1 - G\left(\frac{\sigma_o^2}{\sigma_1^2}\chi^2_{n-1,\alpha/2}\right) \\
&= pchisq\left(\frac{\sigma_o^2}{\sigma_1^2}qchisq(\alpha/2, n-1), n-1\right) + 1 - pchisq\left(\frac{\sigma_o^2}{\sigma_1^2}qchisq(1-\alpha/2, n-1), n-1\right)
\end{aligned}$$

where $G(\cdot)$ is the cdf of a chi-square distribution with $df = n-1$.

The power function for a size $\alpha = .05$ test of $H_o : \sigma = 1.5$ versus $H_1 : \sigma \neq 1.5$ based on a random sample of size $n = 10$ is given on the next page.

Notice that the power curve is definitely not symmetric about $\sigma = 1.4$ as was the case for the power curves for 2-sided alternatives when the test statistic was a t-test or Z-test.


POWER Function For 2−Sided Test of Sigma
POWER
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
H0 σ = 1.5
H1 σ ≠ 1.5
n=10, alpha=.05
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Sigma

Robustness of Chi-squared Test for σ

The inference methods about σ are based on the condition that the random sample is selected from a population having a normal distribution similarly to the requirements for using t-distribution based inference procedures.

However, when sample sizes are moderate to large, the t-distribution based procedures can be used to make inferences about μ even when the normality condition does not hold, since for moderate to large sample sizes the Central Limit Theorem provides that the sampling distribution of the sample mean is approximately normal.

Unfortunately, the same type of result does not hold for the chi-square based procedures for making inferences about σ.

That is, if the population distribution is distinctly nonnormal then the chi-square based inference procedures for σ are not appropriate even if the sample size is large.

Population nonnormality, in the form of skewness or heavy tails, can have serious effects on the nominal significance and coverage probabilities for σ. If a box plot or normal probability plot of the sample data shows substantial skewness or a substantial number of outliers, the chi-squred based inference procedures should not be applied.

When the population distribution is non-normal, there a number of alternative approaches.

One such procedure is the bootstrap. We can use bootstrap techniques to estimate the sampling distribution of sample variance. The estimated sampling distribution is then manipulated to produce confidence intervals for σ and rejection regions for tests of hypotheses about σ.

At the end of this handout, a method will be described which uses bootstrap confidence intervals to test hypotheses.

Simulation Study for Impact of Non-normal Distributions

A simulation study was conducted to investigate the effect on the level of the chi-square test of sampling from heavy tailed and skewed distributions rather than the required normal distribution. The five distributions were normal, uniform (short tailed), t-distribution with df=5 (heavy tailed), and two gamma distributions, one slightly skewed and the other heavily skewed. Summary statistics about the distributions are given in the following table.

Summary Statistics for Distributions in Simulation

	Distribution				
Summary Statistic	Normal	Uniform	t (df=5)	Gamma (shape=1)	Gamma (shape =.1)
Mean	0	17.32	0	10	3.162
Variance	100	100	100	100	100
Skewness	0	0	0	2	6.32
Kurtosis	3	1.8	9	9	63

Note that each of the distributions has the same variance, $\sigma^2 = 100$, but the skewness and kurtosis of the distributions vary. From each of the distributions, 2500 random samples of size 10, 20 and 50 were selected and a test of $H_o : \sigma^2 \leq 100$ vs $H_1 : \sigma^2 > 100$ and a test of $H_o : \sigma^2 \geq 100$ vs $H_1 : \sigma^2 < 100$ were conducted using $\alpha = .05$ for both sets of hypotheses.

A chi-square test of variance was performed for each of the 2500 samples of the various sample sizes from each of the five distributions. The results are given in the following table. The values in the following table are estimates of the maximum probability of Type I error, the significance level, for the chi-square test of hypotheses about variances. If the test statistic was robust to departures from the normal distribution, then all the values in the following table would be very close to 0.05.

Proportion of Times $H_o : \sigma \leq 100$ Was Rejected ($\alpha = .05$)

	Distribution				
Sample Size	Normal	Uniform	t (df=5)	Gamma (shape=1)	Gamma (shape =.1)
n = 10	.046	.018	.119	.202	.213
n = 20	.050	.011	.140	.213	.578
n = 50	.051	.018	.157	.220	.528

Proportion of Times $H_o : \sigma \geq 100$ Was Rejected ($\alpha = .05$)

	Distribution				
Sample Size	Normal	Uniform	t (df=5)	Gamma (shape=1)	Gamma (shape =.1)
n = 10	.047	.004	.083	.134	.139
n = 20	.052	.006	.103	.139	.175
n = 50	.049	.004	.122	.156	.226

The above results are consistent with our observations concerning the sampling distribution of the sample standard deviation, S. That is, the shape of the distribution of S becomes more highly right skewed when sampling from a population which has a right skewed distribution.

1. When the samples are taken from a normal population, the actual probabilities are very nearly equal to the nominal value, $\alpha = 0.05$. The deviation from 0.05 is due to simulation variability.

2. When the population distribution is symmetric with shorter tails than the normal distribution, the actual significance levels are smaller than 0.05

3. For a symmetric distribution with heavy tails, the significance level are much greater than 0.05.

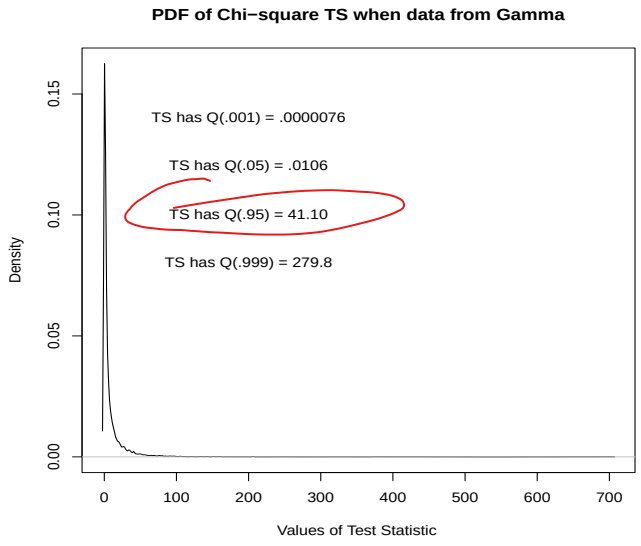
4. For skewed distributions, the actual α values are much larger than the nominal 0.05 value. Furthermore, as the population distribution becomes more skewed, the deviation from 0.05 increases.

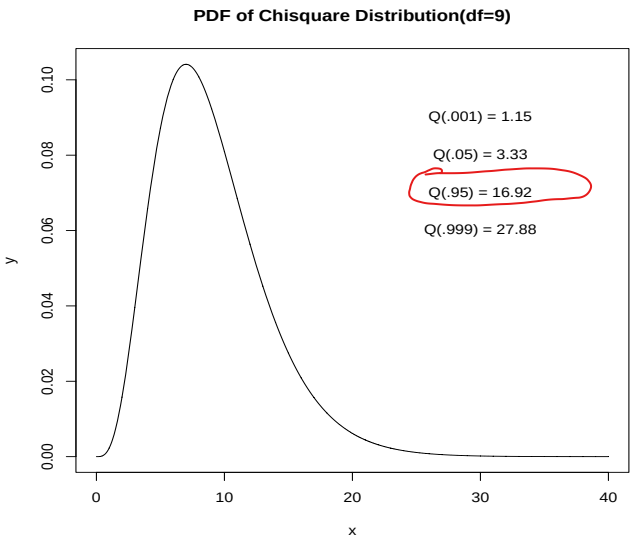
5. From these results, there is strong evidence that the nominal α value of the chi-square test of hypotheses about population variances is very sensitive to non-normality.

6. **This strongly reinforces our recommendation to evaluate the normality of the data prior to conducting the chi-square test of a hypothesis about a population variance.**

7. The article, "Testing Variance of Skewed Distributions", by S.J. Lee and Ping Sa in *Communications in Statistics*, Vol. 27, pp. 807-822, provide a test procedure for the situation where the population distribution is skewed.

To gain further insight, consider the following: I generated 10,000 samples of size 10 from a Gamma distribution with $\alpha = .1$ and $\beta = 31.6$, this is a highly right skewed distribution with var=100. Next compute S^2 from each sample and compute $TS = (n-1)S^2/100$ from each of the 10,000 samples yielding 10,000 values of the TS. From these 10,000 values, I obtained estimates of the upper and lower percentiles of the distribution of TS: the lower .05 percentile was .0106 and the upper .05 percentile was 41.099. If the data had been simulated from a normal distribution with variance=100 then the distribution of TS would be chisquare with df=9 and would have lower .05 percentile equal to 3.325 and upper .05 percentile equal to 16.919. Thus, if we ran the test as if the data was from a normal distribution, then reject $Ho : \sigma^2 = 100$ if TS > 16.919. However, the upper .05 percentile for the TS when the data is from a highly skewed Gamma dist is 41.099 therefore the TS will reject Ho too often, that is, the probability of a Type I error would be much larger than .05. A similar thing happens in the opposite case. Reject $Ho : \sigma^2 = 100$ if $TS < 3.325$. However, the lower .05 percentile for the TS when the data is from a highly skewed Gamma dist is .0106 from the simulation therefore the TS will reject Ho too often, that is, the probability of a Type I error would be much larger than .05. The following graph depicts the pdf of the sampling distribution of the $TS = (n-1)S^2/100$ when sampling from a normal distribution and when sampling from a Gamma($\gamma = .1, \beta = 31.6$) distribution.





$\alpha = P[TS \geq 16.92] > P[TS \geq 41.099] = 0.05$

START Friday 11/12/21

Tests about a Population Proportion - p

Let p be a population/process proportion. For example, the proportion of a population having some characteristic, Type A or the proportion of a process outcomes which are deemed successes.

Let Y be a random variable which describes one of the following situations:

a. Y is the number of Type A units in a random sample of n units selected with replacement from a finite population or

b. Y is the number of Type A units in a random sample of n units selected with or without replacement from an essentially infinite population

c. Y is the number of successes in n iid Bernoulli trials.

Y has a $Bin(n,p)$ distribution and $\hat{p} = \frac{Y}{n}$

We will consider several test statistics for testing hypotheses about p:

Small sample size n: $\min[np_o, n(1-p_o)] < 5$

When the sample size n is small $\min[np_o, n(1-p_o)] < 5$, we will use the $Bin(n,p)$ distribution to set critical values, compute p-values and power values.

Case 1. To test $H_o : p \leq p_o$ versus $H_1 : p > p_o$

Reject H_o if $Y \geq B_{\alpha,p_o} = qbinom(1-\alpha, n, p_o)$, where B_{α,p_o} is the upper α percentile of the $Bin(n,p_o)$ distribution.

$p - value = P[B \geq Y] = 1 - G(Y-1) = 1 - pbinom(Y-1, n, p_o)$,

where $G(\cdot)$ is the cdf of a $Bin(n,p_o)$ distribution.

Power Function: $\rightarrow$ $P[\text{reject } H_0 \text{ when } p = p_1]$ $\rightarrow$ Prob correctly reject null.

$\gamma(p_1) = P_{p_1}[Y \geq B_{\alpha,p_o}] = 1 - G(B_{\alpha,p_o} - 1) = 1 - pbinom(qbinom(1-\alpha, n, p_o) - 1, n, p_1)$,

where $G(\cdot)$ is the cdf of a $Bin(n,p_1)$ distribution.

Case 2. To test $H_o : p \geq p_o$ versus $H_1 : p < p_o$

Reject H_o if $Y \leq B_{1-\alpha,p_o} = qbinom(\alpha, n, p_o)$, where $B_{1-\alpha,p_o}$ is the lower α percentile of the $Bin(n,p_o)$ distribution.

$p - value = P[B \leq Y] = G(Y) = pbinom(Y, n, p_o)$,

Power Function:

$\gamma(p_1) = P_{p_1}[Y \leq B_{1-\alpha,p_o}] = G(B_{1-\alpha,p_o}) = pbinom(qbinom(\alpha, n, p_o), n, p_1)$,

Case 3. To test $H_o : p = p_o$ versus $H_1 : p \neq p_o$

Reject H_o if $Y \leq B_{\alpha/2,p_o} = qbinom(\alpha/2, n, p_o)$ or $Y > B_{1-\alpha/2,p_o} = qbinom(1 - alpha/2, n, p_o)$, where $B_{1-\alpha/2,p_o}$ is the lower $\alpha/2$ percentile and $B_{\alpha/2,p_o}$ is the upper $\alpha/2$ percentile of the $Bin(n, p_o)$ distribution.

$$\begin{aligned} p - value &= 2min[P[B \leq Y],\ P[B \geq Y] \\ &= 2min[G(Y),\ 1 - G(Y-1)] \\ &= 2min[pbinom(Y, n, p_o),\ 1 - pbinom(Y-1, n, p_o)] \end{aligned}$$

Power Function:

$$\begin{aligned} \gamma(p_1) &= P_{p_1}[Y \leq B_{1-\alpha/2,p_o}] + P_{p_1}[Y \geq B_{\alpha/2,p_o}] \\ &= G(B_{1-\alpha/2,p_o}) + 1 - G(B_{\alpha/2,p_o} - 1) \\ &= pbinom(qbinom(\alpha/2, n, p_o), n, p_1) + 1 - pbinom(qbinom(1 - \alpha/2, n, p_o) - 1, n, p_1) \end{aligned}$$

Asymptotic Test Statistic for n large and p_o not too close to 0 or 1

In the case where n is large, we can use the central limit theorem, to state that the sampling distribution of the test statistic when $p = p_o$:

$$TS = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$$

approaches a $N(0, 1)$ distribution as $n \to \infty$ when $p = p_o$.

When n is large, $min[np_o, n(1 - p_o)] \geq 5$, we will use the $N(0, 1)$ distribution to set the critical values and calculate p-values and power values when using the above test statistic:

An alternative approaches, uses the test statistics which was the pivot in construction confidence intervals for p :

$$TS = \frac{\hat{p} - p_o}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

The major problem with using this test statistic is that it is very difficult to calculate the power of the associated test due to the denominator depending of the sample data and not the true value of p. Also, it has the same limitations that we determined in using the above statistics as a pivot in obtaining C.I.'s for p, recall the Wald C.I. for p.

Case 1. To test $H_o : p \leq p_o$ versus $H_1 : p > p_o$

Reject H_o if $TS \geq Z_\alpha$, where $Z_\alpha = qnorm(1-\alpha)$ is the upper α percentile of the $N(0,1)$ distribution.

$p - value = P[Z \geq TS] = 1 - \Phi(TS) = 1 - pnorm(TS)$,

where $\Phi(\cdot)$ is the cdf of a $N(0,1)$ distribution.

Power Function:

$$\begin{aligned}
\gamma(p_1) &= P_{p_1}\left(\frac{\hat{p}-p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} \geq Z_\alpha\right) \\
&= P_{p_1}\left(\hat{p} \geq p_o + Z_\alpha\sqrt{\frac{p_o(1-p_o)}{n}}\right) \\
&= P_{p_1}\left(\hat{p}-p_1 \geq p_o - p_1 + Z_\alpha\sqrt{\frac{p_o(1-p_o)}{n}}\right) \\
&= P_{p_1}\left(\frac{\hat{p}-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} \geq Z_\alpha\sqrt{\frac{p_o(1-p_o)}{p_1(1-p_1)}} + \frac{\sqrt{n}(p_o-p_1)}{\sqrt{p_1(1-p_1)}}\right) \\
&= 1 - \Phi\left(Z_\alpha\sqrt{\frac{p_o(1-p_o)}{p_1(1-p_1)}} + \frac{\sqrt{n}(p_o-p_1)}{\sqrt{p_1(1-p_1)}}\right) \\
&= 1 - pnorm\left(Z_\alpha\sqrt{\frac{p_o(1-p_o)}{p_1(1-p_1)}} + \frac{\sqrt{n}(p_o-p_1)}{\sqrt{p_1(1-p_1)}}\right)
\end{aligned}$$

Case 2. To test $H_o : p \geq p_o$ versus $H_1 : p < p_o$

Reject H_o if $TS \leq -Z_\alpha$, where $Z_\alpha = qnorm(1-\alpha)$ is the upper α percentile of the $N(0,1)$ distribution.

$p - value = P[Z \leq TS] = \Phi(TS) = pnorm(TS)$,

For testing $H_o : p \geq p_o$ versus $H_1 : p < p_o$

Power Function:

$$\begin{aligned}
\gamma(p_1) &= P_{p_1}\left(\frac{\hat{p}-p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} \leq -Z_\alpha\right) \\
&= \Phi\left(-Z_\alpha\sqrt{\frac{p_o(1-p_o)}{p_1(1-p_1)}} + \frac{(p_o-p_1)\sqrt{n}}{\sqrt{p_1(1-p_1)}}\right) \\
&= pnorm\left(-Z_\alpha\sqrt{\frac{p_o(1-p_o)}{p_1(1-p_1)}} + \frac{(p_o-p_1)\sqrt{n}}{\sqrt{p_1(1-p_1)}}\right)
\end{aligned}$$

Case 3. To test $H_o : p = p_o$ versus $H_1 : p \neq p_o$

Reject H_o if $|TS| \geq Z_{\alpha/2}$, where $Z_{\alpha/2} = qnorm(1-\alpha/2)$ is the upper $\alpha/2$ percentile of the $N(0,1)$ distribution.

$p - value = P[Z \geq |TS|] = 2(1 - \Phi(|TS|)) = 2(1 - pnorm(abs(TS))$,

Power Function:

$$\begin{aligned}
\gamma(p_1) &= P_{p_1}\left(\frac{|\hat{p}-p_o|}{\sqrt{\frac{p_o(1-p_o)}{n}}} \geq Z_{\alpha/2}\right) \\
&= \Phi\left(-Z_{\alpha/2}\sqrt{\frac{p_o(1-p_o)}{p_1(1-p_1)}} + \frac{(p_o-p_1)\sqrt{n}}{\sqrt{p_1(1-p_1)}}\right) + 1 - \Phi\left(Z_{\alpha/2}\sqrt{\frac{p_o(1-p_o)}{p_1(1-p_1)}} + \frac{(p_o-p_1)\sqrt{n}}{\sqrt{p_1(1-p_1)}}\right) \\
&= pnorm\left(-Z_{\alpha/2}\sqrt{\frac{p_o(1-p_o)}{p_1(1-p_1)}} + \frac{(p_o-p_1)\sqrt{n}}{\sqrt{p_1(1-p_1)}}\right) + 1 - pnorm\left(Z_{\alpha/2}\sqrt{\frac{p_o(1-p_o)}{p_1(1-p_1)}} + \frac{(p_o-p_1)\sqrt{n}}{\sqrt{p_1(1-p_1)}}\right)
\end{aligned}$$

Sample Size Determination

We can compute sample size n requirements based on specifications placed on the size and power of the test:

Find the minimum n such that a size α test will have $\gamma(p_1) \geq 1-\beta$ whenever $|p_1 - p_o| \geq \delta$:

For $H_1 : p > p_o$, we have

$$1-\beta = \gamma(p_1) = 1 - \Phi\left(Z_\alpha\sqrt{\frac{p_o(1-p_o)}{p_1(1-p_1)}} + \frac{(p_o-p_1)\sqrt{n}}{\sqrt{p_1(1-p_1)}}\right)$$

$$\Phi(-Z_\beta) = \beta \Rightarrow -Z_\beta = Z_\alpha\sqrt{\frac{p_o(1-p_o)}{p_1(1-p_1)}} + \frac{(p_o-p_1)\sqrt{n}}{\sqrt{p_1(1-p_1)}} \Rightarrow$$

$$n = \left[\frac{Z_\alpha\sqrt{p_o(1-p_o)} + Z_\beta\sqrt{p_1(1-p_1)}}{\delta}\right]^2$$

The sample size formula for $H_1 : p < p_o$ is identical to the above. For the two-sided alternative $H_1 : p \neq p_o$, replace Z_α with $Z_{\alpha/2}$

Example We are asked to design a test of $H_o : p \leq .10$ versus $H_1 : p > .10$. The researcher specifies that they want a size .05 test with power of at at least .99 of detecting that the process proportion is .25 or larger.

Solution: Find minimum n such that a size .05 test will have power of at$\gamma(p) \geq .99$ whenever $p \geq .25$, that is, $\delta = p_1 - p_o \geq .25 - .1 = .15$

Take

$$\begin{aligned} n &= \left[\frac{Z_{.05}\sqrt{.1(1-.1)} + Z_{.01}\sqrt{.25(1-.25)}}{.15}\right]^2 \\ &= \left[\frac{1.645\sqrt{.1(1-.1)} + 2.326\sqrt{.25(1-.25)}}{.15}\right]^2 \\ &= 100.09 \end{aligned}$$

Thus, we need at least n=101 observations to met the researchers specifications.

Equivalence Between C.I. and Tests of Hypotheses

We can develop a test of hypotheses using a C.I. for the parameter being tested:

A C.I. $(\hat{\theta}_L,\ (\hat{\theta}_U)$ is a set of feasible values for θ. If θ_o is in C.I. then data supports $H_o : \theta = \theta_o$.

Case 1. Two-sided test of size α for $H_o : \theta = \theta_o$ versus $H_1 : \theta \neq \theta_o$.

An alternative to setting up the rejection region is to construct a

$100(1-\alpha)\%$ C.I. $(\hat{\theta}_L, \hat{\theta}_U)$ for θ and

then use the following rejection region:

Reject H_o if $\theta_o \notin (\hat{\theta}_L, \hat{\theta}_U)$, that is, if $\theta_o \leq \hat{\theta}_L$ or if $\theta_o \geq \hat{\theta}_U \ \Rightarrow$ p-value $\leq \alpha$

The size of this test would be determined by

$$\max_{\theta \in \Theta_o} P[\text{Type I error}] = P_{\theta_o}[RejectH_o]$$

$$\begin{aligned} P_{\theta_o}[RejectH_o] &= P_{\theta_o}[\theta_o \notin (\hat{\theta}_L, \hat{\theta}_U)] \\ &= 1 - P_{\theta_o}[\theta_o \in (\hat{\theta}_L, \hat{\theta}_U)] \\ &= 1 - P[\text{Coverage of CI}] \\ &= 1 - (1-\alpha) = \alpha \end{aligned}$$

This demonstrates that the procedure has the correct size.

In fact the C.I. is identical to the two sided test in those situations where the test statistic and pivot are the same.

Case 2 One-sided test of size α for $H_o : \theta \leq \theta_o$ versus $H_1 : \theta > \theta_o$.

An alternative to setting up the rejection region is to construct a

$100(1-\alpha)\%$ lower bound on θ: $(\hat{\theta}_L, \infty)$,

That is, what is the smallest potential value of θ that the data will support?

Then use the following rejection region:

Reject H_o if $\theta_o \notin (\hat{\theta}_L, \infty)$, that is, Reject H_o if $\hat{\theta}_L > \theta_o \ \Rightarrow \ p - value \leq \alpha$

Thus, reject H_o if the feasible values of θ based on the data, $(\hat{\theta}_L, \infty)$, are all greater than θ_o.

Case 3 One-sided test of size α for $H_o : \theta \geq \theta_o$ versus $H_1 : \theta < \theta_o$.

Construct a $100(1-\alpha)\%$ upper bound on θ: $(-\infty, \hat{\theta}_U)$, then

Reject H_o if $\theta_o \notin (-\infty, \hat{\theta}_U)$, that is, Reject H_o if $\hat{\theta}_U < \theta_o \ \Rightarrow \ p - value \leq \alpha$,

that is, the feasible values of θ are all less than θ_o.

EXAMPLE Rejection Region Based on $100(1-\alpha)\%$C.I. for μ

Suppose we want to test the hypotheses:

$H_o : \mu = \mu_o$ versus $H_1 : \mu \neq \mu_o$

where μ is the mean of a population having a $N(\mu, \sigma^2)$ distribution.

Let $\mathbf{Y} = (Y_1, \cdots, Y_n)$ be a random sample from the population, i.e., $Y_1, \cdots, Y_n$ are iid $N(\mu, \sigma^2)$, we will consider the unrealistic situation where σ is known.

A $100(1-\alpha)\%$ C.I. for μ is :

$$\left(\bar{Y} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{Y} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Therefore, our level α test of H_o vs H_1 would be

Reject H_o if

$$\mu_o \notin \left(\bar{Y} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{Y} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

That is,

Reject H_o when

$$\mu_o < \bar{Y} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ or } \mu_o > \bar{Y} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ or}$$

That is,

Reject H_o when

$$\bar{Y} - \mu_o > Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ or } \bar{Y} - \mu_o < -Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

That is,

Reject H_o when

$$|\bar{Y} - \mu_o| > Z_{1-\alpha/2}\sigma/\sqrt{n}$$

The above is the standard rejection region for testing $H_o : \mu = \mu_o$ versus $H_1 : \mu \neq \mu_o$ when σ is specified.

Bootstrap Tests of Hypotheses

Tests based on Bootstrap C.I.

We can apply this general procedure of testing hypotheses based on a C.I. to implementing bootstrap principles in the testing situation.

Case 1 To test the hypotheses $H_o : \theta = \theta_o$ versus $H_1 : \theta \neq \theta_o$ at level α,

Construct a bootstrap $100(1-\alpha)\%$ C.I. for θ, $(\hat{\theta}_L,\ \hat{\theta}_U)$ and

Reject H_o if $\theta_o \notin (\hat{\theta}_L,\ \hat{\theta}_U)$, that is, if $\theta_o \leq \hat{\theta}_L$ or if $\theta_o \geq \hat{\theta}_U$

P-value: If θ_o is outside the C.I., then we can conclude $p - value \leq \alpha$

If θ_o is within the C.I., then we can conclude $p - value > \alpha$

Case 2 To test the hypotheses $H_o : \theta \leq \theta_o$ versus $H_1 : \theta > \theta_o$ at level α,

Construct a bootstrap $100(1-\alpha)\%$ lower Confidence Bound on θ, $(\hat{\theta}_L, \infty)$ and

Reject H_o if $\theta_o \notin (\hat{\theta}_L, \infty)$, that is, if $\theta_o \leq \hat{\theta}_L$

P-value: If $\theta_o \leq \hat{\theta}_L$., then we can conclude $p - value \leq \alpha$

If $\theta_o > \hat{\theta}_L$, then we can conclude $p - value > \alpha$

Case 3 To test the hypotheses $H_o : \theta \geq \theta_o$ versus $H_1 : \theta < \theta_o$ at level α,

Construct a bootstrap $100(1-\alpha)\%$ upper Confidence Bound on θ, $(-\infty, \hat{\theta}_U)$ and

Reject H_o if $\theta_o \notin (-\infty, \hat{\theta}_U)$, that is, if $\theta_o \geq \hat{\theta}_U$

P-value: If $\theta_o \geq \hat{\theta}_U$, then we can conclude $p - value \leq \alpha$

If $\theta_o < \hat{\theta}_U$, then we can conclude $p - value > \alpha$

Tests based on Bootstrap Test Statistic

In those situations where the hypotheses involve a parameter, θ, (mean, variance, proportion, etc.), then the following procedure can be used to obtain a bootstrap approximation to the p-value.

Let $\hat{\theta}$ be an estimator of θ with estimated standard error, $\hat{SE}(\hat{\theta})$

Let T be the corresponding studentized test statistic, that is,

$$T = \frac{\hat{\theta} - \theta_o}{\hat{SE}(\hat{\theta})}$$

1. From the data, $X_1,\ X_2,\ \ldots,\ X_n$, compute $T = t_o$

2. Transform the data to $Y_i = f(X_i)$ such that Y_i's produce $\hat{\theta}_Y = \theta_o$

   - For example, if testing $H_o : \mu \leq \mu_o$ vs $H_1 : \mu > \mu_o$, where μ is the population mean, then

     $Y_i = X_i - \bar{X} + \mu_o$ which would yield

     $\hat{\mu}_Y = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X} + \mu_o) = \bar{X} - \bar{X} + \mu_o = \mu_o$

   - If testing $H_o : \sigma \leq \sigma_o$ vs $H_1 : \sigma > \sigma_o$, where σ is the population standard deviation, then

     $Y_i = \sigma_0 X_i / S_x$ which would yield

     $\bar{Y} = \sigma_0 \bar{X} / S_x$ and

     $\hat{\sigma}_Y^2 = S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\sigma_0 X_i/S_x - \sigma_0 \bar{X}/S_x)^2 = \frac{\sigma_0^2}{S_x^2}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{\sigma_0^2}{S_x^2}S_x^2 = \sigma_0^2$

3. Construct B bootstrap samples of size n by sampling with replacement from $Y_1, Y_2, \ldots, Y_n$.

4. Compute a value of T from each of the B bootstrap samples: $t_1^*,\ t_2^*,\ \ldots,\ t_B^*$

5. For each of the following cases, we obtain an estimated p-value for test H_o versus H_1:

   Case 1 To test the hypotheses $H_o : \theta = \theta_o$ versus $H_1 : \theta \neq \theta_o$

   $p - value \approx [\ \#\text{samples with } |t_i^*| \geq |t_o|\ ]/B$

   Case 2 To test the hypotheses $H_o : \theta \leq \theta_o$ versus $H_1 : \theta > \theta_o$

   $p - value \approx [\ \#\text{samples with } t_i^* \geq t_o\ ]/B$

   Case 3 To test the hypotheses $H_o : \theta \geq \theta_o$ versus $H_1 : \theta < \theta_o$

   $p - value \approx [\ \#\text{samples with } t_i^* \leq t_o\ ]/B$

EXAMPLE Suppose we have a random sample $X_1,\ X_2,\ \ldots,\ X_n$ from a population with mean μ and we want to test

$$H_o:\ \mu \leq \mu_o \ \text{ versus } \ H_1 : \mu > \mu_o$$

If the population was approximately normally distributed with mean μ then the appropriate test statistics would be

$$T = \frac{\hat{\theta} - \theta}{\hat{SE}(\hat{\theta})} = \frac{\bar{X} - \mu_o}{S/\sqrt{n}}$$

However, if the population distribution is highly right skewed then the computation of the p-value using a t-distribution with $df\ =\ n-1$ would be very inaccurate.

A bootstrap procedure could be applied as an alternative to using the Wilcoxon signed rank test or the sign test.

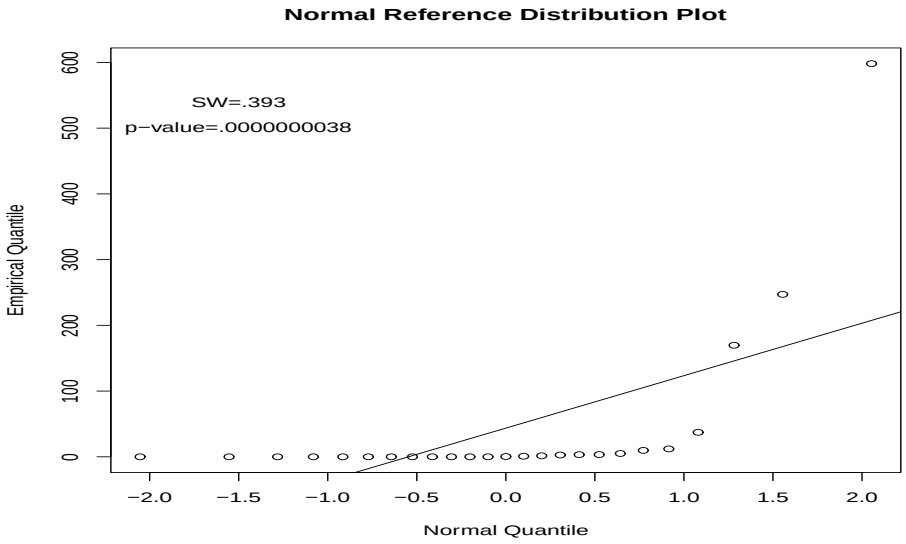
Let $Y_i = X_i - \bar{X} + \mu_o$ and then generate the bootstrap samples from the Y_i's

An excellent reference which discusses some of the problems you may encounter in applying a bootstrap test procedure is **Bootstrap Methods and Their Applications** by A.C. Davison and D.V. Hinkley.

EXAMPLE A researcher wants to test if the population standard deviation is greater than 100. He takes a random sample of 25 values from the population:

0.0	0.0	0.0000068	0.0000135	0.0000176	0.0000209	0.0003327	0.007412	0.0147
0.03630	0.08643	0.08937	0.4875	0.8276	1.406	2.685	3.294	
3.391	4.989	9.712	12.01	37.29	169.7	247.1	598.1	

The data produces a standard deviation of 129.36 which would seem to indicate that σ is greater than 100, that is, $H_1 : \sigma > 100$. Based on the normal reference plot and the Shapiro-Wilk test there is strong evidence that the data is not normally distributed.



The researcher ignores the non-normality and calculates a lower bound for σ and a p-value for the test using the normal based procedure:

$\frac{(n-1)S^2}{\sigma_o^2} = \frac{(25-1)(129.36)^2}{(100)^2} = 40.16 \;\Rightarrow\; p - value = 1 - pchisq(40.16, 24) = .021$

The normal based lower 95% confidence bound for σ is $\sqrt{\frac{(n-1)S^2}{\chi^2(.05,24)}} = 105.02$ which is greater than 100 and hence would support the conclusion that σ is greater than 100.

The company's statistician informs the researcher that it would be a good idea to use a bootstrap procedure because the test statistics is highly sensitive to a lack of normality. The following R code produces a bootstrap a p-value for testing $H_o : \sigma \leq 100$ versus $H_1 : \sigma > 100$. Recall that the p-value is the probability of obtaining a value of the test statistic larger than the value of the test statistic from the data set under the assumption that H_o is true, that is, $\sigma = 100$. Thus, it is necessary to transform the data into a data set having a standard deviation of 100 prior to conducting the resampling. Transform the data, X's, by $Y = 100 * X/sd(X)$ which yields a data set having $sd(Y) = 100$

The following R code (bootTestofSigma.R) yields the bootstrap p-value and a bootstrap lower 95% confidence bound on σ:

STOP CN

Friday 11/12/21

NOTE: Check for corrected version of 1.1.0.12 w/ following code changed to make it more readable.

```
x = c(0.0,0.0,0.0000068,0.0000135,0.0000176,0.0000209,0.0003327,0.007412,
      0.0147,0.03630,0.08643,0.08937,0.4875,0.8276,1.406,2.685,3.294,3.391,4.989,
      9.712,12.01,37.29,169.7,247.1,598.1)

n = length(x)
x = sort(x)
i = seq(1:n)
u = (i-.5)/n
z = qnorm(u)
plot(z,x,datax=F,plot=T,xlab="Normal Quantile",ylab="Empirical Quantile",
       lab=c(7,8,7),
       main="Normal Reference Distribution Plot",
       cex=.95)
abline(lm(x~z))
shapiro.test(x)
text(-1.5,540,"SW=.393")
text(-1.5,500,"p-value=.0000000038")
m = mean(x)
SD = sd(x)
sigma0 = 100
TSD = (n-1)*SD^2/sigma0^2
B  =  9999
PV =  numeric(B)
PV =  rep(0,B)
TS =  numeric(B)
TS =  rep(0,B)
S =  numeric(B)
S =  rep(0,B)
Sy =  numeric(B)
Sy =  rep(0,B)

{
for (i in 1:B)
S[i]  =  sd(sample(x,replace=T))
}

PV = (n-1)*S^2/SD^2
PV =  sort(PV)

#standardize the data so that the null hypothesis is true

y = x*sigma0/sd(x)

{
for (i in 1:B)
Sy[i]  =  sd(sample(y,n,replace=T))
}
```

```
TS = (n-1)*Sy^2/sigma0^2

SIMpvalue = sum(TS>TSD)/B

NORMpvalue = 1-pchisq(TSD,n-1)

LPV  =  PV[250]
UPV  =  PV[9750]

LPV2 = PV[9500]

SIMci = c(sqrt((n-1)*SD^2/UPV), sqrt((n-1)*SD^2/LPV))

SIMLcb = sqrt((n-1)*SD^2/LPV2)

NORMLcb = sqrt((n-1)*SD^2/qchisq(.95,n-1))



SD
# 129.3631

SIMpvalue
# .2022

NORMpvalue
# .0202

SIMLcb
# 84.15

NORMLcb
105.02
```

Note that the bootstrap lower 95% confidence bound is 84.15 which is less than 100 and the bootstrap p-value is .2022 both of which imply that there is not significant evidence that σ is greater than 100.

C.I. Based on a Test of Hypotheses

We can develop a C.I. for a parameter using a test of hypotheses for the parameter:

We want to construct a $100(1-\alpha)\%$ C.I. for a parameter θ where the feasible values of θ are in the parameter space Θ.

Suppose for each $\eta \in \Theta$ we have a level α test of

$$H_o : \theta = \eta \ \text{ vs } \ H_1 : \theta \neq \eta$$

For data vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, define $C(\mathbf{y})$ by

$$C(\mathbf{y}) = \{\eta : H_o : \theta = \eta \ \text{ is not rejected at level } \alpha\}$$

.

Then, a $100(1-\alpha)\%$ confidence interval for θ is $C(\mathbf{Y})$.

That is,

$$P_\theta\left(\theta \in C(\mathbf{Y})\right) \geq 1-\alpha \ \text{ for all } \ \theta \in \Theta$$

EXAMPLE

Suppose we want a $100(1-\alpha)\%$ C.I. for σ^2 where the population distribution is $N(\mu,\ \sigma^2)$.
A level α test of $H_o : \sigma = \sigma_o$ versus $H_1 : \sigma \neq \sigma_o$ is

Reject H_o if $\frac{(n-1)S^2}{\sigma_o^2} \leq \chi^2_{n-1,1-\alpha/2}$ or $\frac{(n-1)S^2}{\sigma_o^2} \geq \chi^2_{n-1,\alpha/2}$

That is, Do not Reject H_o if $\chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma_o^2} \leq \chi^2_{n-1,\alpha/2}$
Therefore, a $100(1-\alpha)\%$ C.I. for σ^2 is

$$\begin{aligned} C(\mathbf{y}) &= \{\sigma_o : H_o : \sigma = \sigma_o \ \text{ is not rejected at level } \alpha\} \\ &= \left\{\sigma_o : \chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma_o^2} \leq \chi^2_{n-1,\alpha/2}\right\} \\ &= \left\{\sigma_o : \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma_o^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right\} \end{aligned}$$

Therefore, the $100(1-\alpha)\%$ C.I. for σ^2 is

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}},\ \ \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right)$$

Test of whether autocorrelation function is equal to 0

For a stationary times series: Y_t; $t = 1, \ldots, n$, with n "large", the doted lines on the acf graph are a level $\alpha = .05$ test of whether or not the population autocorrelations $\rho_k = Corr(Y_t, Y_{t-k})$ are equal to 0 or not. The dotted lines are placed at

$$-1.96/\sqrt{n} \text{ and } 1.96/\sqrt{n}$$

In a plot of 40 or so values of $\hat{\rho}_k$, if more than two or three values of $\hat{\rho}_k$ fall outside the dotted, then we would conclude that there is significant evidence that the random variables Y_t are not independent but have nonzero autocorrelations. In the acf for the Stamford Ozone data, we have $n = 136$ so the dotted lines are places at

$$-1.96/\sqrt{n} \text{ and } 1.96/\sqrt{n} = -1.96/\sqrt{136} \text{ and } 1.96/\sqrt{136} = (-.168, .168)$$

From the plot, we have that the values of $\hat{\rho}_1$, $\hat{\rho}_2$ and $\hat{\rho}_6$ are outside the bounds. Therefore, we can conclude there is evidence of correlation in the ozone data for Stamford.