# Stat 427/527: Advanced Data Analysis I

Chapter 2: Estimation in One-Sample Problems

Instructor: Yan Lu

THE UNIVERSITY of
NEW MEXICO

# Topics

- Inference for a population mean.
- Confidence intervals.
- Hypothesis testing.
- Statistical versus practical significance.
- Design issues and power.

## Overview

- ▶ Identify a population of interest
  —-for example, UNM freshmen female students' weight, height or entrance GPA.

- ▶ Population parameters
  —-unknown quantities of the population that are of interest, say, population mean $\mu$ and population variance $\sigma^2$ etc.

- ▶ Random sample
  —-Select a random or representative sample from the population.
  —-A sample consists random variables $Y_1, \cdots, Y_n$, that follows a specified distribution, say $N(\mu, \sigma^2)$

- ▶ Statistic: a function of radom variables $Y_1, \ldots, Y_n$, which does not depend on any unknown parameters

- ▶ Observed sample: $y_1, y_2, \cdots, y_n$ are observed sample values after data collection

▶ We cannot see much of the population
— but would like to know what is typical in the population
— The only information we have is that in the sample.

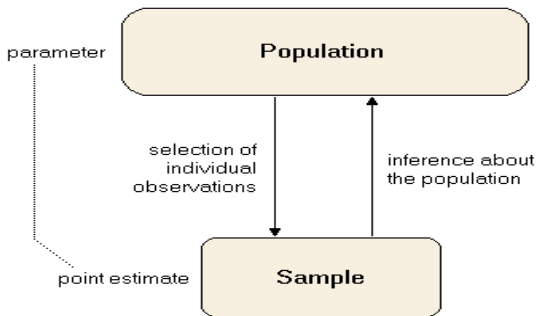**Goal**: want to use the sample information to make inferences about the population and its parameters.



Figure 1 : Population, sample and statistical inference

Notations:

- Population mean: $\mu$
- Sample mean: $\bar{Y} = \sum_{i=1}^{n} Y_i/n$
- Estimate of sample mean: the value of $\bar{Y}$ computed from data $\bar{y} = \sum_{i=1}^{n} y_i/n$
- Population variance: $\sigma^2$
- Sample variance: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
- Estimate of sample variance: the value of $S^2$ computed from data $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$
- Population standard deviation: $\sigma$
- Sample standard deviation (Standard error): $S$
- Estimate of standard error: $s$, the value of $S$ computed from data

Table 1 : Commonly seen parameters, statistics and estimates:

| Parameters | Statistic | Estimate |
|---|---|---|
| Describe a popn | Describe a random sample | Describe an observed sample |
| $\mu$ | $\bar{Y}$ | $\bar{y}$ |
| $\sigma^2$ | $S^2$ | $s^2$ |
| $\sigma$ | $S$ | $s$ |

## 2.1 Inference for a population mean

Notations:

- ▶ Parameter of interest: population mean $\mu$
- ▶ Sample mean: $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}$.
- ▶ Observed sample mean: $\bar{y} = \sum_{i=1}^{n} y_i / n$

Two main methods for inferences on $\mu$:

- ▶ **Confidence intervals** (CI)
- ▶ **Hypothesis tests**

# Sampling distribution

Sampling distribution: probability distribution of a given statistic
based on a random sample
—-Statistic is also a r.v.
—-Sampling distribution is in contrast to the population
distribution
Want to know the sampling distribution of $\bar{Y}$
Recall that

- ▶ standard error (SE): the standard deviation of the sampling
  distribution of a statistic
- ▶ Standard error of the mean (SEM): is the standard deviation
  of the sample-mean's estimator

If $Y_1, \ldots, Y_n$ are observations of a random sample of size $n$ from normal distributions $N(\mu, \sigma^2)$ and $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ is the sample mean of the $n$ observations. We have

$$SE_{\bar{Y}} \;\; = \;\; s/\sqrt{n}$$

where

$s$ is the sample standard deviation (i.e., the sample-based estimate of the standard deviation of the population)

$n$ is the size (number of observations) of the sample.

## Central limit theorem (CLT)

If $Y_1, \ldots, Y_n$ is a random sample of size $n$ taken from a population or a distribution with mean $\mu$ and variance $\sigma^2$ and if $\bar{Y}$ is the sample mean, then for large $n$,

$$\bar{Y} \sim N(\mu, \sigma^2/n)$$

## illustration of CLT

- ▶ Consider random variables $Y_i \sim$ Uniform$(0, 1)$ distribution
  —- any value in the interval $[0, 1]$ is equally likely
  —- $\mu = E(Y) = 1/2$, and $\sigma^2 = Var(Y) = 1/12$, so the standard deviation is $\sigma = \sqrt{1/12} = 0.289$.
- ▶ Draw a sample of size $n$
  —- the standard error of the mean will be $\sigma/\sqrt{n}$
  —- as $n$ gets larger the distribution of the mean will increasingly follow a normal distribution.
  **Illustration:**
1. generate unifrom random sample of size $n$
2. calculate sample mean $\bar{y}$
3. repeat for $N = 10000$ times
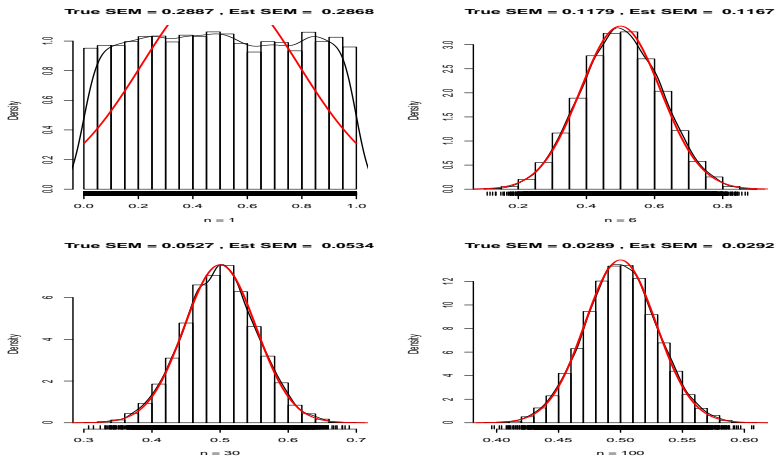4. plot those $N$ means, compute the estimated SEM

Figure 2 : illustration of CLT, notice even with samples as small as 2 and 6 that the properties of the SEM and the distribution are as predicted

## illustration of CLT

In a more extreme example, we draw samples from an Exponential(1) distribution ($\mu = 1$ and $\sigma = 1$), which is strongly skewed to the right.

$$f(x) = e^{-x}, x > 0$$

Notice that the normality promised by the CLT requires larger samples sizes, about $n \geq 30$, than for the previous Uniform(0,1) example, which required about $n \geq 6$.
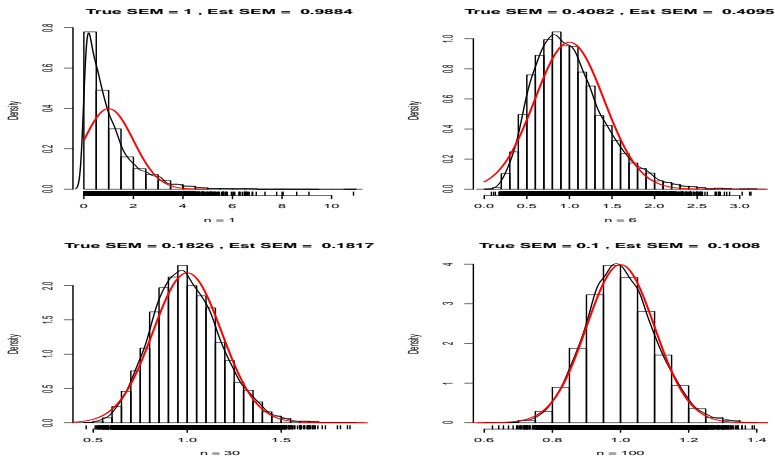
Figure 3 : illustration of CLT, notice that the normality promised by the CLT requires larger samples sizes, about $n \geq 30$

Note that the further the population distribution is from being normal, the larger the sample size is required to be for the sampling distribution of the sample mean to be normal.

Question: If the population distribution is normal, what's the minimum sample size for the sampling distribution of the mean to be normal?

## Standardization

If $Y_1, \ldots, Y_n$ is a random sample of size $n$ taken from a normal population with mean $\mu$ and variance $\sigma^2$ and if $\bar{Y}$ is the sample mean, then,

$$\bar{Y} \sim N(\mu, \sigma^2/n).$$

We may standardize $\bar{Y}$ by subtracting the mean and dividing by the standard deviation, which results in the variable

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

and

$$Z \sim N(0, 1)$$

## $t$-distribution

The Student's $t$-**distribution** is a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the *sample size is small* and population *standard deviation is unknown*.

▶ $t$-distribution is symmetric and bell-shaped, like the normal distribution, but has heavier tails, meaning that it is more prone to producing values that fall far from its mean.

▶ the $t$-distribution is wider than the normal distribution because in addition to estimating the mean $\mu$ with $\bar{Y}$, we *also* have to estimate $\sigma^2$ with $S^2$, so there's some additional uncertainty.

▶ The degrees-of-freedom (df) parameter of the $t$-distribution is the sample size $n$ minus the number of variance parameters estimated. Thus, $df = n - 1$ when we have one sample and $df = n - 2$ when we have two samples.

▶ As $n$ increases, the $t$-distribution becomes close to the normal distribution, and when $n = \infty$ the distributions are equivalent.
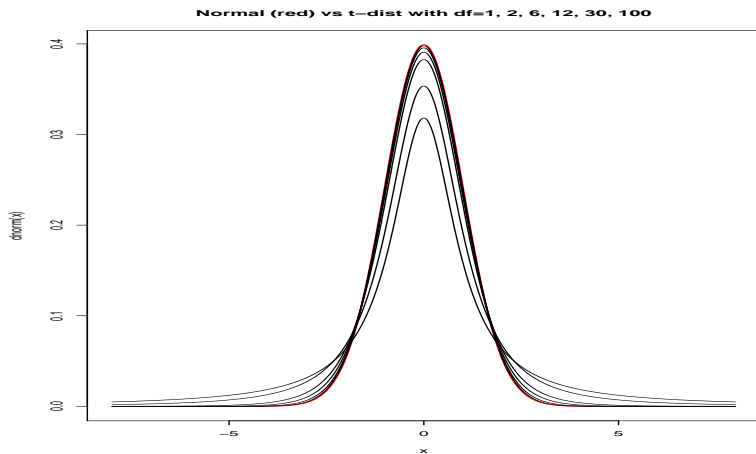
Figure 4 : Normal (red) vs t-distributions with a range of degrees-of-freedom df=1, 2, 6, 12, 30, 100

# Confidence Interval (CI) for $\mu$, variance unknown

Confidence interval: an interval estimate $[l, u]$ for a population parameter, say, $\mu$.

—— a range of plausible values for $\mu$, with $l$ the lower bound, and $u$ the upper bound, based on the observed data

——Best Guess $\pm$ Reasonable Error of the Guess.

▶ If $Y_1, Y_2, \ldots, Y_n$ is a random sample from normal distribution with mean $\mu$ and variance $\sigma^2$ ($\sigma^2$ is unknown), i.e. $Y_i \stackrel{iid}{\sim} N(\mu, \sigma^2), i = 1, \ldots, n$. The r.v.

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}}$$

has a $t$ distribution with $n - 1$ degrees of freedom.

▶ Confidence coefficient $\alpha$: a number between 0 and 100%.

—- $t_{\alpha/2}$ is a number such that $p(T \leq t_{\alpha/2}) = 1 - \alpha/2$. The number $t_{\alpha/2}$ is often called upper $100\alpha/2$ percentage point of the $t$ distribution.



**t PDF (Two-Sided Test at Alpha = 0.05)**
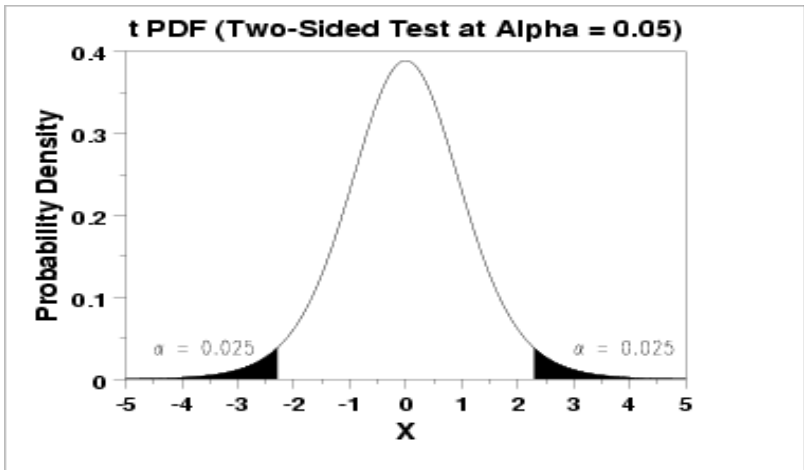
▶ Further, We can show that

$$P(-t_{\alpha/2} \leq \frac{\bar{Y} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2}) = 1 - \alpha$$

$$P(-t_{\alpha/2}S/\sqrt{n} + \mu \leq \bar{Y} \leq t_{\alpha/2}S/\sqrt{n} + \mu) = 1 - \alpha, \text{ equivalently}$$

$$P\left(\bar{Y} - t_{\alpha/2} * \frac{S}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{\alpha/2} * \frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

# The $t$ Confidence Interval on $\mu$

If $\bar{y}$ is the sample mean of an observed sample $(y_1, \ldots, y_n)$ from a normal population with unknown variance $\sigma^2$ and unknown mean $\mu$, then a $100(1 - \alpha)\%$ CI on $\mu$ is given by

$$[\bar{y} - t_{\alpha/2, n-1} * \frac{s}{\sqrt{n}}, \bar{y} + t_{\alpha/2, n-1} * \frac{s}{\sqrt{n}}]$$

▶ *Interpretation*: the observed interval $[l, u]$, contains the true value of $\mu$ (interpret $\mu$ in the context, for example, mean income level), with confidence $100(1 - \alpha)\%$.

▶ If you repeatedly sample the population and construct 95% CIs for $\mu$, then 95% of the intervals will contain $\mu$, whereas 5% will not.

Recall a $100(1 - \alpha)\%$ CI on $\mu$ is given by

$$[\bar{y} - t_{\alpha/2,n-1} * \frac{s}{\sqrt{n}}, \bar{y} + t_{\alpha/2,n-1} * \frac{s}{\sqrt{n}}]$$

Notes that the length of the interval estimate is $2 * t_{\alpha/2} * s/\sqrt{n}$, then

- As $\alpha \uparrow$, the confidence $(1 - \alpha)\%$, $t_{\alpha/2} \downarrow$, and hence the confidence interval gets narrower.
- As $s \uparrow$, the confidence interval gets wider.
- As $n \uparrow$, the confidence interval gets narrower.

# An example with 100 CIs

- Consider drawing a sample of 25 observations from a normally distributed population with mean 10 and sd 2.
- Calculate the 95% $t$-CI.
- Repeat that 100 times.

The plot belows reflects the variability of that process. We expect 95 of the 100 CIs to contain the true population mean of 10, that is, on average 5 times out of 100 we draw the incorrect inference that the population mean is in an interval when it does not contain the true value of 10.
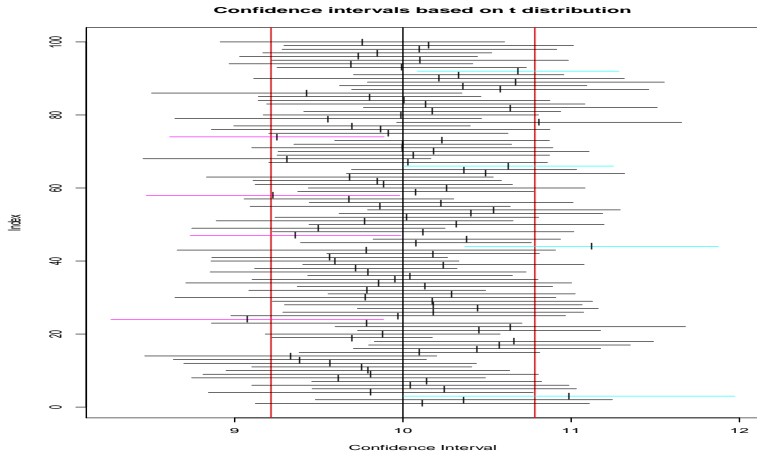
Figure 6 : green and red intervals didn't contain true mean 10

## Assumptions for the t CI procedures

Recall t-distribution If $Y_1, Y_2, \ldots, Y_n$ is a random sample from normal distribution with mean $\mu$ and variance $\sigma^2$ ($\sigma^2$ is unknown), i.e. $Y_i \stackrel{iid}{\sim} N(\mu, \sigma^2), i = 1, \ldots, n$. The r.v.

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}}$$

has a $t$ distribution with $n-1$ degrees of freedom.

- Data are a **random sample** from the population of interest
- **Population frequency curve is normal**

   —- The normality assumption can never be completely verified without having the entire population data.

   —-You can assess the reasonableness of this assumption using a stem-and-leaf display, a boxplot and a histogram of the sample data if we assume that sample is representative of the population.

   —-The stem-and-leaf and histogram display from the data should resemble a normal curve.

- In fact, the assumptions are slightly looser than this, the population frequency curve can be anything provided the sample size is large enough that it's reasonable to assume that the **sampling distribution of the mean is normal**.
- In the cases that sample size is small and the distribution of the sample data is not normal, we can use bootstrap procedure to check for the normality of the sampling distribution of the mean

## Example: Age at First Heart Transplant

Let us go through a hand-calculation of a CI, and also use R to generate summary data. We are interested in the mean age at first heart transplant for a population of patients.

> **1. Define the population parameter, plot the data**
>
> Let $\mu$ = mean age at the time of first heart transplant for a population of patients.

```
#### Example: Age at First Transplant
# enter data as a vector
age <- c(54, 42, 51, 54, 49, 56, 33, 58, 54, 64, 49)
>summary(age)
Min. 1st Qu.  Median   Mean 3rd Qu.   Max.
33.00   49.00   54.00   51.27   55.00   64.00

> # stem-and-leaf plot
> stem(age, scale=2)
  The decimal point is 1 digit(s) to the right of the |

  3 | 3
  3 |
  4 | 2
  4 | 99
  5 | 1444
  5 | 68
  6 | 4
```
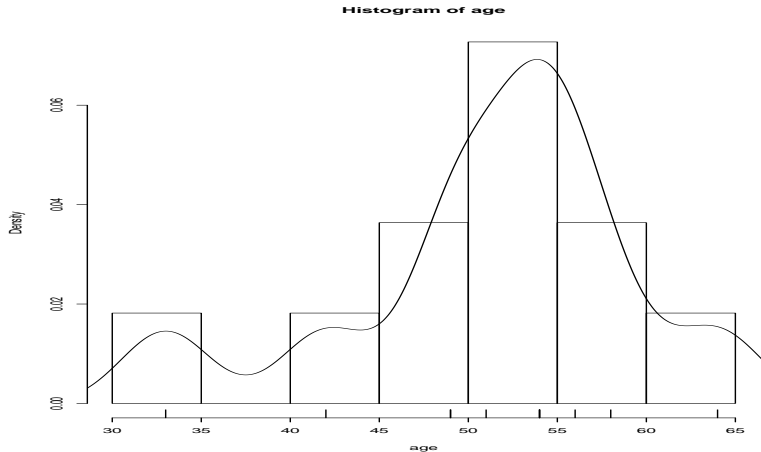
Figure 7 : Histogram plot of age

▶ **2. Calculate summary statistics from sample**
The ages (in years) at first transplant for a sample of 11 heart transplant patients are as follows:

54, 42, 51, 54, 49, 56, 33, 58, 54, 64, 49.

Summaries for the data are: $n = 11$, $\bar{y} = 51.27$, and $s = 8.26$ so that $SE_{\bar{Y}} = 8.26/\sqrt{11} = 2.4904$. The degrees of freedom are $df = 11 - 1 = 10$, and $t_{\mathrm{crit}} = t_{0.025} = 2.228$.

*Now calculate the confidence interval by hand.*

### 3. Specify confidence level, find critical value, calculate limits

Let us calculate a 95% CI for $\mu$. For a 95% CI $\alpha = 0.05$, so we need to find $t_{\text{crit}} = t_{0.025}$, which is 2.228. Now $t_{\text{crit}} \times SE_{\bar{Y}} = 2.228 \times 2.4904 = 5.55$. The lower limit on the CI is $l = 51.27 - 5.55 = 45.72$. The upper limit is $u = 51.27 + 5.55 = 56.82$.

```
> # t.crit
> qt(1 - 0.05/2, df = length(age) - 1)
[1] 2.228139
```

### 4. Summarize in words
For example, I am 95% confident that the mean age at first transplant of a population is between 45.7 and 56.8 years (rounding off to 1 decimal place).
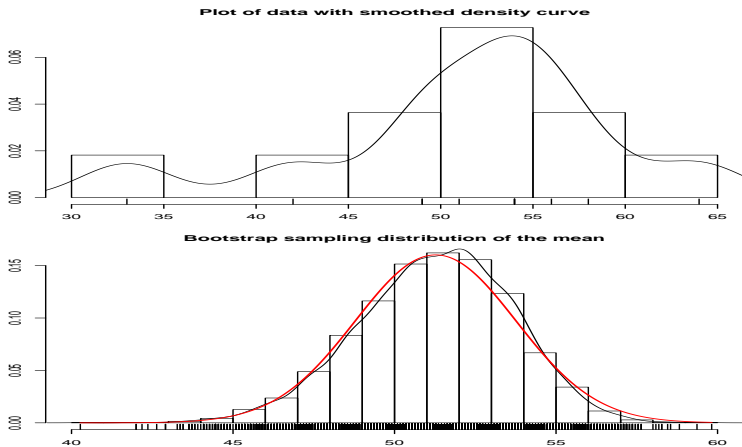
▶ **5. Check assumptions**



Figure 8 : Plot of data with smoothed density curve and bootstrap sampling distribution of the mean

In the cases that sample size is small and the distribution of the sample data is not normal, we can use bootstrap procedure to check for the normality of the sampling distribution of the mean.

The assumption of normality of the sampling distribution of the mean appears reasonablly close. In fact, if the data is not extremely skewed or with extreme outliers, $t$ approximation of the mean is appropriate. Therefore, the results for the $t$ confidence interval above can be trusted.

▶ **6. Now do the calculation in R by yourself**

## Statistical hypothesis:

- ▶ Statistical hypothesis is a statement about the parameters of one or more populations.
- ▶ Because we use probability distributions to represent populations, a statistical hypothesis may also be thought of as a statement about the probability distribution of a random variable.

Examples:

- ▶ a) The chance of showing up head in tossing a coin is 0.5, i.e. $p = 0.5$, or the chance is not 0.5, i.e. $p \neq 0.5$.
- ▶ b) The average age of first year college student is 18, i.e. $\mu = 18$, or the average age is greater than 18, i.e. $\mu > 18$.

**A hypothesis test** often consider two competing hypotheses.

- ▶ One hypothesis is called null hypothesis, denoted as $H_0$.
- ▶ The other hypothesis is called the alternative hypothesis, denoted as $H_1$ or $H_\alpha$.

Let $\theta$ be a parameter of a population and $\theta_0, \theta_1$ are two specific real values. The following gives a summary of the possible combination we are interested in.

- ▶ Two sided alternative hypothesis:
  a) $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$.
- ▶ One sided alternative hypothesis:
  b) $H_0 : \theta = \theta_0$, $H_1 : \theta < \theta_0$.
  c) $H_0 : \theta = \theta_0$, $H_1 : \theta > \theta_0$.

## Test of hypothesis

**Test of a hypothesis**: a procedure leading to a decision about the null hypothesis.

—-We take a random sample and see which of the two hypotheses our data is most consistent with. If data information is consistent with the null hypothesis, we will not reject it; if this information is inconsistent with the null hypothesis, we will reject the null hypothesis and in favor of the alternative.

—-A test statistic is a single measure of some attribute of a sample (i.e. a statistic) used in statistical hypothesis testing. In different hypothesis testing problems, different test statistics are used. Let $h(Y_1, \ldots, Y_n)$ denote the test statistic. Sample test statistic is then $h(y_1, \ldots, y_n)$

## Type I and II errors:

Consider the three scenarios of hypothesis testing

a) $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$.

b) $H_0 : \theta = \theta_0$, $H_1 : \theta < \theta_0$, (c) $H_0 : \theta = \theta_0$, $H_1 : \theta > \theta_0$.

**Acceptance region**: a region $[l, u]$ for which we will fail to reject the null hypothesis when the **sample test statistic** is in the region. The boundaries of the acceptance region are called **critical values**.

**Rejection region**: a region for which we reject the null hypothesis when the test statistic is in the region. The rejection region is the complementary region of the acceptance region.
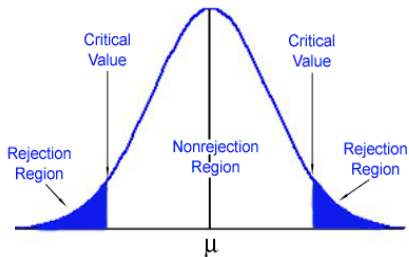
**Type I error**: rejecting the null hypothesis $H_0$ when it is true.
**Type II error** : failing to reject the null hypothesis when it is false.
**Probability of Type I error**:

$$\alpha = P(\text{reject } H_0 \text{ when } H_0 \text{ is true }) = P(h(Y_1, \ldots, Y_n) \notin [l, u])|\theta = \theta_0)$$

Probability of Type I error is also called significance level, or size of the test.
**Probability of Type II error**

$$\begin{aligned} \beta &= P(\text{ fail to reject } H_0 \text{ when } H_0 \text{ is false }) \\ &= P(h(Y_1, \ldots, Y_n) \in [l, u])|\theta = \theta_1) \end{aligned}$$

where $\theta_1$ is the true population parameter value.
**Power of a statistical test**: the probability of rejecting the null hypothesis $H_0$ when the alternative hypothesis is true. It is computed as $1 - \beta$.

|  | State of nature | |
|---|---|---|
| **Decision** | $H_0$ true | $H_A$ true |
| Fail to reject $H_0$ | correct decision | *Type-II error* |
| Reject $H_0$ in favor of $H_A$ | *Type-I error* | correct decision |

P-value: the P-value is the probability of obtaining a test statistic result at least as extreme as the one that was actually observed, assuming that the null hypothesis and all the other assumptions used in the test is true. Smaller P-value indicates greater evidence against the null hypothesis or $H_0$ is less plausible.

- a) $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$.

  $$\text{P-value} = P\left(|h(X_1, \ldots, X_n)| > |h(x_1, \ldots, x_n)| \mid \theta = \theta_0\right)$$

- b) $H_0 : \theta = \theta_0$, $H_1 : \theta < \theta_0$.

  $$\text{P-value} = P(h(X_1, \ldots, X_n) < h(x_1, \ldots, x_n)|\theta = \theta_0)$$

- c) $H_0 : \theta = \theta_0$, $H_1 : \theta > \theta_0$.

  $$\text{P-value} = P(h(X_1, \ldots, X_n) > h(x_1, \ldots, x_n)|\theta = \theta_0)$$
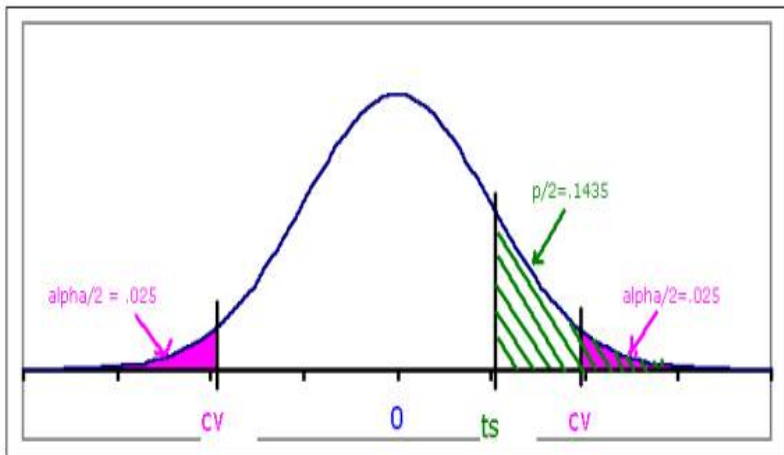
Figure 10 : Consider a two-sided test, green shaded area is $1/2$ of the pvalue, red shaded area is 0.05 corresponding to the critical value (CV)

# Tests on the mean of a normal distribution, variance unknown

The test statistic is

$$T_0 = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}.$$

Hypothesis of testing $H_0 : \mu = \mu_0$ vs the following alternative hypotheses are summarized in the table.

Table 2 : $\bar{y}$ is sample mean and $s$ is sample standard deviation; $t_{n-1,\alpha/2}$ is the upper $\alpha/2$ percentage points of the t distribution with $n-1$ degrees of freedom; $t_{n-1,\alpha}$ is the upper $\alpha$ percentage points of the t distribution with $n-1$ degrees of freedom; $T_{n-1}$ is a random variable following t distribution with $n-1$ degrees of freedom. $\alpha$ is the significance level of the test

| Step 1: | $H_1 : \mu \neq \mu_0$ |
|---------|-------------------------|
| Step 2: | compute $t_0 = \frac{\bar{y}-\mu_0}{s/\sqrt{n}}$ |
| Step 3a: | Reject $H_0$ if $t_0 > t_{n-1,\alpha/2}$ or $t_0 < -t_{n-1,\alpha/2}$ |
| Step 3b: | P-value $= 2P(T_{n-1} > |t_0|)$ |
|  | Reject $H_0$ if P-value $< \alpha$ |
| Power | $P(T_0 > t_{n-1,\alpha/2}|\mu_1) + P(T_0 < -t_{n-1,\alpha/2}|\mu_1)$ |

Table 3 :  $\bar{y}$ is sample mean and $s$ is sample standard deviation; $t_{n-1,\alpha/2}$ is the upper $\alpha/2$ percentage points of the t distribution with $n-1$ degrees of freedom; $t_{n-1,\alpha}$ is the upper $\alpha$ percentage points of the t distribution with $n-1$ degrees of freedom; $T_{n-1}$ is a random variable following t distribution with $n-1$ degrees of freedom. $\alpha$ is the significance level of the test

| Step 1: | $H_1 : \mu < \mu_0$ | $H_1 : \mu > \mu_0$ |
|---------|---------------------|---------------------|
| Step 2: | compute $t_0 = \frac{\bar{y}-\mu_0}{s/\sqrt{n}}$ | compute $t_0 = \frac{\bar{y}-\mu_0}{s/\sqrt{n}}$ |
| Step 3a: | Reject $H_0$ if $t_0 < -t_{n-1,\alpha}$ | Reject $H_0$ if $t_0 > t_{n-1,\alpha}$ |
| Step 3b: | P-value $= P(T_{n-1} < t_0)$ | P-value $= P(T_{n-1} > t_0)$ |
|         | Reject $H_0$ if P-value $< \alpha$ | Reject $H_0$ if P-value $< \alpha$ |
| Power | $P(T_0 < -t_{n-1,\alpha/2}|\mu_1)$ | $P(T_0 > t_{n-1,\alpha/2}|\mu_1)$ |

Exercise: Consider the scores of stat 427 students last year.
Suppose we randomly choose 15 students out of the entire group
of stat 427 students and denote $\bar{Y}$ as the sample mean. Assume
that the sample standard deviation $s = 3$. Suppose our acceptance
region for testing $H_0 : \mu = 80$ is $[-2, 2]$ and the test Statistic is
$\frac{\bar{Y}-80}{S/\sqrt{n}}$, i.e. we fail to reject the null if $\frac{\bar{y}-80}{s/\sqrt{n}} \in [-2, 2]$. What would
be our probability of Type I error? If the true population is normal
with mean score 75. What would be our probability of Type II
error?
Exercise continued. Suppose the class score average is 78, what
would be the P-value for testing $H_0 : \mu = 80$?

# Example: Age at First Transplant (Revisited)

The ages (in years) at first transplant for a sample of 11 heart transplant patients are as follows: 54, 42, 51, 54, 49, 56, 33, 58, 54, 64, 49. Summaries for these data are:

$$n = 11, \bar{y} = 51.27, s = 8.26 \text{ and } SE_{\bar{Y}} = 2.4904.$$

Test the hypothesis that the mean age at first transplant is 50, Use $\alpha = 0.05$.

*Solution:* Define

$\mu = $ mean age at time of first transplant for a population of patients.

We are interested in testing

$$H_0 : \mu = 50 \text{ against } H_A : \mu \neq 50.$$

The degrees of freedom are $df = 11 - 1 = 10$. The critical value for a 5% test is $t_{\mathrm{crit}} = t_{0.025} = 2.228$. (Note $\alpha/2 = 0.05/2 = 0.025$).

For the test,

$$t_s = \frac{\bar{y} - \mu_0}{SE_{\bar{Y}}} = \frac{51.27 - 50}{2.4904} = 0.51.$$

Since $t_{\text{crit}} = 2.228$, we do not reject $H_0$ using a 5% test.

▶ Equivalently, the p-value for the test is 0.62, thus we fail to reject $H_0$ because $0.62 > 0.05 = \alpha$. The results of the hypothesis test should not be surprising, since the CI $[45.72, 56.82]$ tells you that 50 is a plausible value for the population mean age at transplant.

▶ the data *could have* come from a distribution with a mean of 50 — this is not convincing evidence that $\mu$ actually *is* 50.

```
> # look at help for t.test
> help(t.test)
>   # defaults include: alternative = "two.sided",
conf.level = 0.95
> t.summary <- t.test(age, mu = 50)
> t.summary


        One Sample t-test

data:  age
t = 0.51107, df = 10, p-value = 0.6204
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 45.72397 56.82149
sample estimates:
mean of x
 51.27273
```

We can also find p-value by

```
pt(observed t statistic, df)
#note this is the probability of the T statistics
#integrated from -infinity to the observed t statistic
> pt(0.51107,10)
[1] 0.6898024
> 1-pt(0.51107,10)
[1] 0.3101976  # this is the right side p-value
> pt(-0.51107,10)
[1] 0.3101976  # this is the left side p-value
```

Therefore, the p-value of the two sided test is
 2*0.31 =0.62

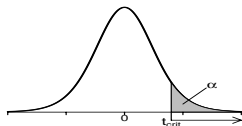## One sided test example:

Table 4 :   Recall one-sided test

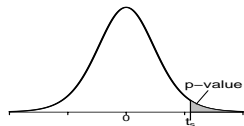| Step 1: | $H_1 : \mu < \mu_0$ | $H_1 : \mu > \mu_0$ |
|---|---|---|
| Step 2: | compute $t_0 = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$ | compute $t_0 = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$ |
| Step 3a: | Reject $H_0$ if $t_0 < -t_{n-1,\alpha}$ | Reject $H_0$ if $t_0 > t_{n-1,\alpha}$ |
| Step 3b: | P-value $= P(T_{n-1} < t_0)$ | P-value $= P(T_{n-1} > t_0)$ |
|  | Reject $H_0$ if P-value $< \alpha$ | Reject $H_0$ if P-value $< \alpha$ |
| Power | $P(T_0 < -t_{n-1,\alpha/2} | \mu_1)$ | $P(T_0 > t_{n-1,\alpha/2} | \mu_1)$ |

# Recall: one sided test
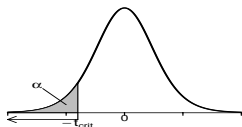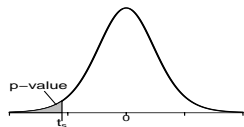
| ■ | CLICKER $Q$s  —  One-sided tests on $\mu$ | ■ |

Example: Weights of canned tomatoes  A consumer group suspects

# Example: Weights of canned tomatoes

A consumer group suspects that the average weight of canned
tomatoes being produced by a large cannery is less than the
advertised weight of 20 ounces.

- the group purchases 14 cans of the canner's tomatoes from
  various grocery stores.
- The weights of the contents of the cans to the nearest half
  ounce were recorded as follows: 20.5, 18.5, 20.0, 19.5, 19.5,
  21.0, 17.5, 22.5, 20.0, 19.5, 18.5, 20.0, 18.0, 20.5.
- Do the data confirm the group's suspicions? Test at the 5%
  level.

- Let $\mu = $ the population mean weight for advertised 20 ounce cans of tomatoes produced by the cannery.
- The company claims that $\mu = 20$.
- The consumer group believes that $\mu < 20$

  $$H_0 : \mu = 20 \text{ against } H_\alpha : \mu < 20.$$

- The consumer group will reject $H_0$ only if the data overwhelmingly suggest that $H_0$ is false.

1. assess the normality assumption prior to performing the *t*-test.



**Histogram of tomato**
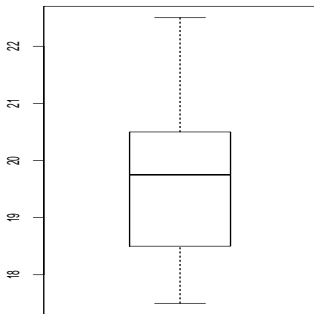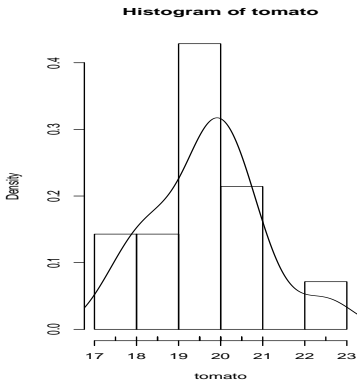
The histogram and the boxplot suggest that the distribution might be slightly skewed to the left. However, the skewness is not severe and no outliers are present, so the normality assumption is not unreasonable.

2. Summary statistics

▶ The sample size, mean, and standard deviation are

$$n = 14, \bar{y} = 19.679, \text{ and } s = 1.295$$

The standard error is $SE_{\bar{Y}} = s/\sqrt{n} = 0.346$.

▶ Sample mean is less than 20. But is it sufficiently less than 20 for us to be willing to publicly refute the canner's claim?

▶ Let us do a hand calculation using the summarized data.
   —- first using the rejection region approach
   —- and then by evaluating a p-value.
   —-find CI.

The test statistic is

$$t_s = \frac{\bar{y} - \mu_0}{SE_{\bar{Y}}} = \frac{19.679 - 20}{0.346} = -0.93.$$

The critical value for a 5% one-sided test is $t_{0.05} = 1.771$

```
> qt(1 - 0.05, df = length(tomato) - 1)
[1] 1.770933
```

▶ reject $H_0$ if $t_s < -1.771$, In our case, $-0.93 > -1.771$, the test statistic is not in the rejection region.

▶ The exact p-value from R is 0.185, which exceeds 0.05.
—Both approaches lead to the conclusion that we do not have sufficient evidence to reject $H_0$.

▶ CI, $(-\infty, 19.679 + 1.77 \times 0.346] = (-\infty, 20.29]$
——As expected, this interval covers 20. That is,

▶ we do not have sufficient evidence to question the accuracy of the canner's claim.

▶ We are 95% cofident that the population mean weight of the canner's 20oz cans of tomatoes is less than or equal to 20.29oz.

```
> t.summary <- t.test(tomato, mu = 20,
 alternative = "less")
> t.summary

        One Sample t-test

data:  tomato
t = -0.92866, df = 13, p-value = 0.185
alternative hypothesis: true mean is less than 20
95 percent confidence interval:
     -Inf 20.29153
sample estimates:
mean of x
 19.67857
```

# Statistical versus practical significance

▶ Statistical significance($\alpha$, p-value): simply mean that the null hypothesis was rejected at the selected significance level.
—-Reflects the odds that a particular finding could have occurred by chance. If the p-value for a difference between two groups is 0.05, it would be expected to occur by chance just 5 times out of 100 (thus, it is likely to be a "real" difference).
—-A small p-value, which would ordinarily indicate statistical significance, may be the result of a large sample size in combination with a departure from $H_0$ that has little practical significance.
—-In many experimental situations, only departures from $H_0$ of large magnitude would be worthy of detection, whereas a small departure from $H_0$ would have little practical significance.

▶ Practical significance
——Reflects the magnitude, or size, of the difference, not the odds that it could have occurred by chance. Arguably much more important than statistical significance, especially for clinical questions.

## Example

Let $\mu$ denote the true average IQ of all children in the very large city of Euphoria. Consider testing

$$H_0 : \mu = 100 \text{ versus } H_\alpha : \mu > 100$$

where $\mu$ is the mean from a normal population with $\sigma = 15$.

- ▶ For a reasonably large sample size $n$, suppose $\bar{y} = 101$ was observed. But one IQ point is no big deal. We would not want this sample evidence to argue strongly for rejection of $H_0$.

- ▶ For various sample sizes, Table (5) records both the P-value when $\bar{y} = 101$ and also the probability of not rejecting $H_0$ at level .01 when $\mu = 101(\beta)$.

Table 5 :   An illustration of the Effect of Sample Size on P-values and type II error $\beta$

| $n$ | P-value when $\bar{y} = 101$ | $\beta(101)$ for level 0.01 test |
|--------|-------------------------|----------------------------------|
| 25     | 0.3085                  | 0.9664                           |
| 100    | 0.1587                  | 0.9082                           |
| 400    | 0.0228                  | 0.6293                           |
| 900    | 0.0013                  | 0.2514                           |
| 1600   | 0.0000335               | 0.0475                           |
| 2500   | 0.000000297             | 0.0038                           |
| 10,000 | $7.69 \times 10^{-24}$  | 0.0000                           |

▶ The second column in Table (5) shows that even for moderately large sample sizes, the P-value indicate strong rejection of $H_0$, whereas the observed $\bar{y}$ itself suggests that in practical terms the true value of $\mu$ differs little from the null value $\mu_0 = 100$.

▶ The third column points out that even when there is little practical difference between the true $\mu$ and the null value, for a fixed level of significance, a large sample size will almost always lead to rejection of the null hypothesis at that level.

▶ One must be especially careful in interpreting evidence when the sample size is large, since any small departure from $H_0$ will almost surely be detected by a test, yet such a departure may have little practical significance.

## Design issues

Sample size for specified error on mean, variance known:

- $\bar{y}$ is an estimate of $\mu$
- we can be $100(1 - \alpha)\%$ confident that the **absolute error** $|\bar{y} - \mu|$ will not exceed a specified amount $E$ when the sample size needed is

$$\left| Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right| \leq E$$

or

$$n \geq \left( \frac{z_{\alpha/2}\sigma}{E} \right)^2.$$

Estimate $\sigma^2$:

- ▶ Pilot sample: a small sample taken to provide information and guidance for the future data collection can be used to estimate quantities needed for setting the sample size
- ▶ Use previous studies or data available in the literature
- ▶ Guess the variance

An experiment may not be sensitive enough to pick up true differences.

- ► Tocopilla meteorite example, suppose the true mean cooling rate is $\mu = 1.00$.
- ► To have a 50% chance of correctly rejecting $H_0 : \mu = 0.54$, you would need about $n = 48$ observations.
- ► If the true mean is $\mu = 0.75$, you would need about 221 observations to have a 50% chance of correctly rejecting $H_0$.
- ► In general, the smaller the difference between the true and hypothesized mean (relative to the spread in the population), the more data that is needed to reject $H_0$.
- ► If you have prior information on the expected difference between the true and hypothesized mean, you can design an experiment appropriately by choosing the sample size required to likely reject $H_0$.