STAT 401A - Statistical Methods for Research Workers

Inference Using t-Distributions

Jarad Niemi (Dr. J)

Iowa State University

last updated: September 8, 2014

Random variables

From: http://www.stats.gla.ac.uk/steps/glossary/probability_distributions.html

Definition

A random variable is a function that associates a unique numerical value with every outcome of an experiment.

Definition

A discrete random variable is one which may take on only a countable number of distinct values such as 0, 1, 2, 3, 4,... Discrete random variables are usually (but not necessarily) counts.

Definition

A continuous random variable is one which takes an infinite number of possible values. Continuous random variables are usually measurements.

Random variables

Examples:

- Discrete random variables
  - Coin toss: Heads (1) or Tails (0)
  - Die roll: 1, 2, 3, 4, 5, or 6
  - Number of Ovenbirds at a 10-minute point count
  - RNAseq feature count
- Continuous random variables
  - Pig average daily (weight) gain
  - Corn yield per acre

Statistical notation

Let Y be 1 if the coin toss is heads and 0 if tails, then

$$Y \sim Bin(n, p)$$

which means

> *Y is a binomial random variable with n trials and probability of success p*

For example, if Y is the number of heads observed when tossing a fair coin ten times, then $Y \sim Bin(10, 0.5)$.

Later we will be constructing $100(1-\alpha)\%$ confidence intervals, these intervals are constructed such that if n of them are constructed then $Y \sim Bin(n, 1-\alpha)$ will cover the true value.

Statistical notation

Let Y_i be the average daily (weight) gain in pounds for the ith pig, then

$$Y_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$$

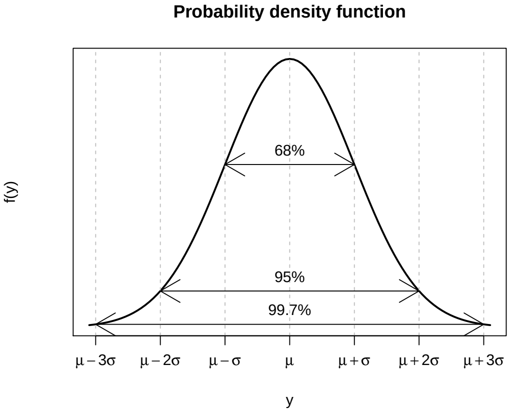
which means

> *Y_i are independent and identically distributed normal (Gaussian) random variables with expected value $E[Y_i] = \mu$ and variance $V[Y_i] = \sigma^2$ (standard deviation σ).*

For example, if a litter of pigs is expected to gain 2 lbs/day with a standard deviation of 0.5 lbs/day and *the knowledge of how much one pig gained does not affect what we think about how much the others have gained*, then $Y_i \stackrel{iid}{\sim} N(2, 0.5^2)$.

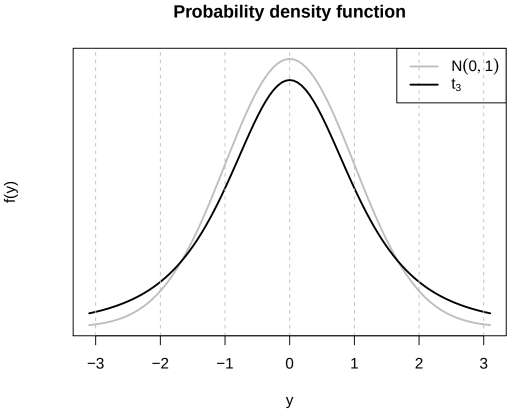
Normal (Gaussian) distribution

A random variable Y has a normal distribution, i.e. $Y \sim N(\mu, \sigma^2)$, with mean μ and variance σ^2 if draws from this distribution follow a bell curve centered at μ with spread determined by σ^2:



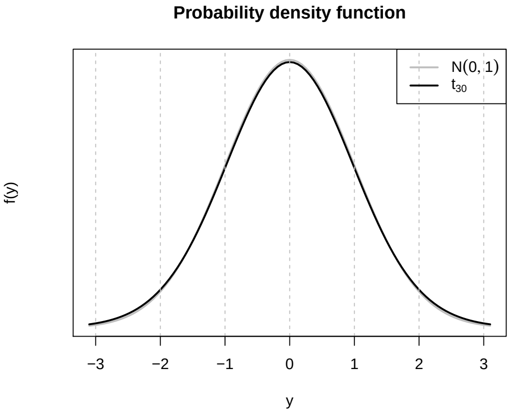
t-distribution

A random variable Y has a t-distribution, i.e. $Y \sim t_v$, with degrees of freedom v if draws from this distribution follow a similar bell shaped pattern:



t-distribution

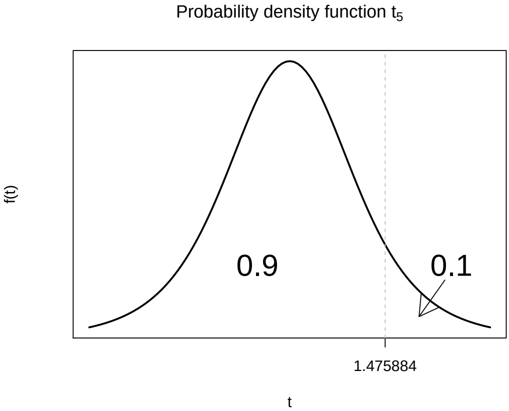
As $v \to \infty$, then $t_v \xrightarrow{d} N(0,1)$, i.e. as the degrees of freedom increase, a t distribution gets closer and closer to a standard normal distribution, i.e. $N(0,1)$. If $v > 30$, the differences is negligible.



t critical value

Definition

If $T \sim t_v$, a $t_v(1-\alpha/2)$ critical value is the value such that $P(T < t_v(1-\alpha/2)) = 1-\alpha/2$ (or $P(T > t_v(1-\alpha)) = \alpha/2$).



Cedar-apple rust

Cedar-apple rust is a (non-fatal) disease that affects apple trees. Its most obvious symptom is rust-colored spots on apple leaves. Red cedar trees are the immediate source of the fungus that infects the apple trees. If you could remove all red cedar trees within a few miles of the orchard, you should eliminate the problem. In the first year of this experiment the number of affected leaves on 8 trees was counted; the following winter all red cedar trees within 100 yards of the orchard were removed and the following year the same trees were examined for affected leaves.

- Statistical hypothesis:
  - H_0: Removing red cedar trees increases or maintains the same mean number of rusty leaves.
  - H_1: Removing red cedar trees decreases the mean number of rusty leaves.

- Statistical question:
  What is the expected reduction of rusty leaves in our sample between year 1 and year 2 (perhaps due to removal of red cedar trees)?

Data

Here are the data

```
library(plyr)
y1 = c(38,10,84,36,50,35,73,48)
y2 = c(32,16,57,28,55,12,61,29)
leaves  = data.frame(year1=y1, year2=y2, diff=y1-y2)
leaves
```

```
  year1 year2 diff
1    38    32    6
2    10    16   -6
3    84    57   27
4    36    28    8
5    50    55   -5
6    35    12   23
7    73    61   12
8    48    29   19
```

```
summarize(leaves, n=length(diff), mean=mean(diff), sd=sd(diff))
```

```
  n mean   sd
1 8 10.5 12.2
```

Is this a statistically significant difference?

Assumptions

Let

- Y_{1j} be the number of rusty leaves on tree j in year 1
- Y_{2j} be the number of rusty leaves on tree j in year 2

Assume

$$D_j = Y_{1j} - Y_{2j} \stackrel{iid}{\sim} N(\mu, \sigma^2)$$

Then the statistical hypothesis test is

H_0: $\mu = 0$ $(\mu \leq 0)$

H_1: $\mu > 0$

while the statistical question is 'what is μ?'

Paired t-test pvalue

Test statistic

$$t = \frac{\overline{D} - \mu}{SE(\overline{D})}$$

where $SE(\overline{D}) = s/\sqrt{n}$ with

- n being the number of observations (differences),
- s being the sample standard deviation of the differences, and
- $\overline{D}$ being the average difference.

If H_0 is true, then $\mu = 0$ and $t \sim t_{n-1}$. The pvalue is $P(t_{n-1} > t)$ since this is a one-sided test. By symmetry, $P(t_{n-1} > t) = P(t_{n-1} < -t)$.

For these data,

$$\overline{D} = 10.5, \mathrm{SE}(\overline{D}) = 4.31, t_7 = 2.43, and \mathrm{p} = 0.02$$

Confidence interval for μ

The 100(1-α)% confidence interval has lower endpoint

$$\overline{D} - t_{n-1}(1-\alpha)SE(\overline{D})$$

and upper endpoint at infinity

For these data at 95% confidence, $t_7(0.9) = 1.89$ and thus the lower endpoint is

$$10.5 - 1.89 \cdot 4.31 = 2.33$$

So we are 95% confident that the true difference in the number of rusty leaves is greater than 2.33.

SAS code for paired t-test

```
DATA leaves;
  INPUT tree year1 year2;
  DATALINES;
1 38 32
2 10 16
3 84 57
4 36 28
5 50 55
6 35 12
7 73 61
8 48 29
;

PROC TTEST DATA=leaves SIDES=U;
    PAIRED year1*year2;
    RUN;
```

SAS output for paired t-test

```
                          The TTEST Procedure

                      Difference:  year1 - year2

 N        Mean      Std Dev      Std Err      Minimum      Maximum

 8     10.5000      12.2007       4.3136      -6.0000      27.0000

    Mean        95% CL Mean          Std Dev       95% CL Std Dev

 10.5000       2.3275   Infty        12.2007       8.0668  24.8317

                          df     t Value     Pr > t

                           7        2.43     0.0226
```

R output for paired t-test

```
t.test(leaves$year1, leaves$year2, paired=TRUE, alternative="greater")



Paired t-test

data:  leaves$year1 and leaves$year2
t = 2.434, df = 7, p-value = 0.02257
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 2.328   Inf
sample estimates:
mean of the differences
                   10.5
```

Statistical Conclusion

Removal of red cedar trees within 100 yards is associated with a significant reduction in rusty apple leaves (paired t-test t_7=2.43, p=0.023). The mean reduction in rust color leaves is 10.5 [95% CI (2.33,∞)].

Do Japanese cars get better mileage than American cars?

- Statistical hypothesis:

  H_0: Mean mpg of Japanese cars is the same as mean mpg of American cars.

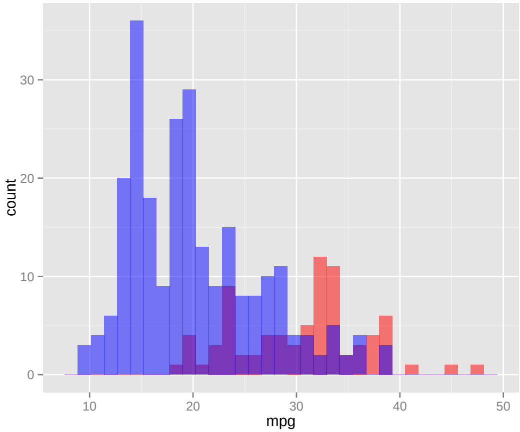
  H_1: Mean mpg of Japanese cars is different than mean mpg of American cars.

- Statistical question:

  What is the difference in mean mpg between Japanese and American cars?

- Data collection:
  - Collect a random sample of Japanese/American cars

```
mpg = read.csv("mpg.csv")
library(ggplot2)
ggplot(mpg, aes(x=mpg))+
  geom_histogram(data=subset(mpg,country=="Japan"), fill="red", alpha=0.5)+
  geom_histogram(data=subset(mpg,country=="US"), fill="blue", alpha=0.5)
```



Assumptions

Let

- Y_{1j} represent the jth Japanese car
- Y_{2j} represent the jth American car

Assume

$$Y_{1j} \stackrel{iid}{\sim} N(\mu_1, \sigma^2) \qquad Y_{2j} \stackrel{iid}{\sim} N(\mu_2, \sigma^2)$$

Restate the hypotheses using this notation

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

Alternatively

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 \neq 0$

Test statistic

The test statistic we use here is

$$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{SE(\overline{Y}_1 - \overline{Y}_2)}$$

where

- $\overline{Y}_1$ is the sample average mpg of the Japanese cars
- $\overline{Y}_2$ is the sample average mpg of the American cars

and

$$SE(\overline{Y}_1 - \overline{Y}_2) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}}$$

where

- s_1 is the sample standard deviation of the mpg of the Japanese cars
- s_2 is the sample standard deviation of the mpg of the American cars

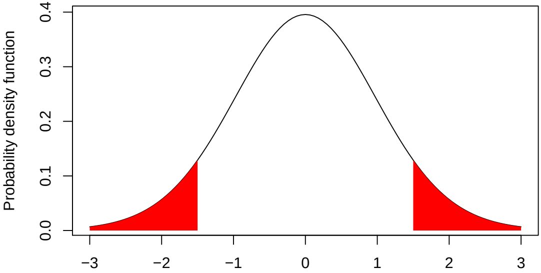
Pvalue

If H_0 is true, then $\mu_1 = \mu_2$ and the test statistic

$$t = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{SE(\overline{Y}_1 - \overline{Y}_2)} \sim t_{n_1+n_2-2}$$

where t_ν is a t-distribution with ν degrees of freedom.

Pvalue is $P(|t_{n_1+n_2-2}| > |t|) = P(t_{n_1+n_2-2} > |t|) + P(t_{n_1+n_2-2} < -|t|)$ or as a picture



Hand calculation

To calculate the quantity by hand, we need 6 numbers:

```
library(plyr)
ddply(mpg, .(country), summarize, n=length(mpg), mean=mean(mpg), sd=sd(mpg))
```

```
  country   n  mean    sd
1   Japan  79 30.48 6.108
2      US 249 20.14 6.415
```

Calculate

$$
\begin{aligned}
s_p &= \sqrt{\frac{(79-1)\cdot 6.11^2+(249-1)\cdot 6.41^2}{79+249-2}} &&= 6.34 \\
SE(\overline{Y}_1 - \overline{Y}_2) &= 6.34\sqrt{\frac{1}{79}+\frac{1}{249}} &&= 0.82 \\
t &= \frac{30.5-20.1}{0.82} &&= 12.6
\end{aligned}
$$

Finally, we are interested in finding $P(|t_{326}| > |12.6|) = 2P(t_{326} < -|12.6|) < 0.0001$ which is found using a table or software.

Confidence interval

Alternatively, we can construct a 100(1-α)% confidence interval. The formula is

$$\overline{Y}_1 - \overline{Y}_2 \pm t_{n_1+n_2-2}(1-\alpha/2) SE(\overline{Y}_1 - \overline{Y}_2)$$

where $\pm$ indicates plus and minus and $t_\nu(1-\alpha/2)$ is the value such that $P(t_\nu < t_\nu(1-\alpha/2)) = 1-\alpha/2$. If $\alpha = 0.05$ and $\nu = 326$, then $t_\nu(1-\alpha/2) = 1.97$.

The 95% confidence interval is

$$30.5 - 20.1 \pm 1.97 \cdot 0.82 = (8.73, 11.9)$$

We are 95% confident that, on average, Japanese cars get between 8.73 and 11.9 more mpg than American cars.

SAS code for two-sample t-test

```
DATA mpg;
    INFILE 'mpg.csv' DELIMITER=',' FIRSTOBS=2;
    INPUT mpg country $;

PROC TTEST DATA=mpg;
    CLASS country;
    VAR mpg;
    RUN;
```

```
                          The TTEST Procedure

                            Variable:  mpg

    country          N        Mean     Std Dev     Std Err     Minimum     Maximum

    Japan           79     30.4810      6.1077      0.6872     18.0000     47.0000
    US             249     20.1446      6.4147      0.4065      9.0000     39.0000
    Diff (1-2)             10.3364      6.3426      0.8190

country       Method               Mean       95% CL Mean        Std Dev      95% CL Std Dev

Japan                           30.4810     29.1130  31.8491      6.1077      5.2814   7.2429
US                              20.1446     19.3439  20.9452      6.4147      5.8964   7.0336
Diff (1-2)    Pooled            10.3364      8.7252  11.9477      6.3426      5.8909   6.8699
Diff (1-2)    Satterthwaite     10.3364      8.7576  11.9152

               Method           Variances        df    t Value    Pr > |t|

               Pooled           Equal           326      12.62      <.0001
               Satterthwaite    Unequal      136.87      12.95      <.0001

                         Equality of Variances

                   Method      Num df    Den df    F Value    Pr > F

                   Folded F       248        78       1.10    0.6194
```

R code/output for two-sample t-test

```
t.test(mpg~country, data=mpg, var.equal=TRUE)



Two Sample t-test

data:  mpg by country
t = 12.62, df = 326, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  8.725 11.948
sample estimates:
mean in group Japan    mean in group US
              30.48               20.14
```

Conclusion

Mean miles per gallon of Japanese cars is significantly different than mean miles per gallon of American cars (two-sample t-test t=12.62, $p < 0.0001$). Japanese cars get an average of 10.3 [95% CI (8.7,11.9)] more miles per gallon than American cars.

Hypotheses

Three key features:

- a test statistic calculated from data
- a sampling distribution for the test statistic under the null hypothesis
- a region that is as or more extreme (one-sided vs two-sided hypotheses)

Calculate probability of being in the region:

Definition

A pvalue is the probability of observing a test statistic as or more extreme than that observed, if the null hypothesis is true.

- If pvalue is less than or equal to α, we reject the null hypothesis.
- If pvalue is greater than α, we fail to reject the null hypothesis.

Hypothesis framework

Let's assume, we have

- $Y_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$ for $i = 1, \ldots, n$ and have
- calculated a test statistic t, and
- if the null hypothesis is true, t has a t_ν sampling distribution.

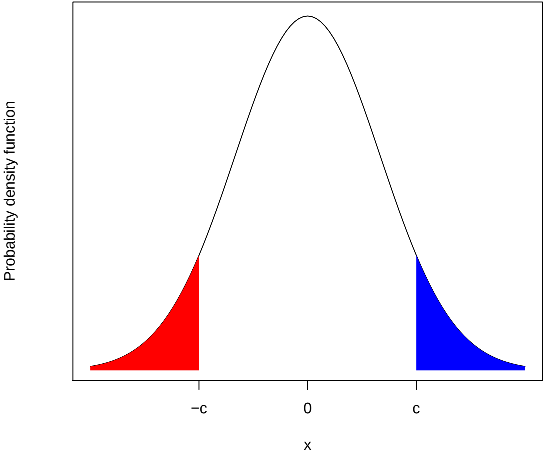
Now, we can have one of three types of hypotheses:

- Two-sided $(H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0)$:
$$\text{pvalue} = P(|t_\nu| > |t|) = P(t_\nu > |t|) + P(t_\nu < -|t|) = 2P(t_\nu < -|t|)$$
- One-sided $(H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0)$:
$$\text{pvalue} = P(t_\nu > t) = P(t_\nu < -t)$$
- One-sided $(H_0 : \mu \geq \mu_0$ vs $H_1 : \mu < \mu_0)$:
$$\text{pvalue} = P(t_\nu < t)$$

$F(c) = P(t_\nu < c)$ is the cumulative distribution function for a t distribution with ν degrees of freedom.

Symmetric distributions

The standard normal and t distributions are both symmetric around zero.



$$P(T_\nu > c) = P(t_\nu < -c) \quad \text{blue area is equal to red area}$$

Paired t-test example

In the paired t-test example, we had a test statistic $t = 2.43$ with a t_7 sampling distribution if the null hypothesis is true.

Consider the following hypotheses (μ is the expected difference):

- Two-sided ($H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$):
$$\text{pvalue} = 2P(t_7 < -2.43) = 0.0454$$
- One-sided ($H_0 : \mu \leq 0$ vs $H_1 : \mu > 0$):
$$\text{pvalue} = \quad P(t_7 < -2.43) = 0.0227$$
- One-sided ($H_0 : \mu \geq 0$ vs $H_1 : \mu < 0$):
$$\text{pvalue} = \quad P(t_7 < 2.43) = 0.9773$$

Two-sample t-test example

In a two-sample t-test, we might have a test statistic $t = -2$ with a t_{30} sampling distribution if the null hypothesis is true.

Consider the following hypotheses ($\mu_1 - \mu_2$ is the expected difference):

- Two-sided ($H_0 : \mu_1 - \mu_2 = 0$ vs $H_1 : \mu_1 - \mu_2 \neq 0$):

$$\text{pvalue} = 2P(t_{30} < -2) = 0.0546$$

- One-sided ($H_0 : \mu_1 - \mu_2 \leq 0$ vs $H_1 : \mu_1 - \mu_2 > 0$):

$$\text{pvalue} = \quad P(t_{30} < 2) = 0.9727$$

- One-sided ($H_0 : \mu_1 - \mu_2 \geq 0$ vs $H_1 : \mu_1 - \mu_2 < 0$):

$$\text{pvalue} = \quad P(t_{30} < -2) = 0.0273$$

Confidence interval construction

Key steps in confidence interval construction:

1. Calculate point estimate
2. Calculate standard error of the statistic
3. Set error level α
4. Find the appropriate critical value
5. Construct the $100(1-\alpha)\%$ confidence interval
    - Two-sided ($H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$): (L, U)

$$(L, U) = \text{estimate} \pm \text{critical value}(1-\alpha/2) \times \text{standard error}$$

    - One-sided ($H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$): (L, ∞)

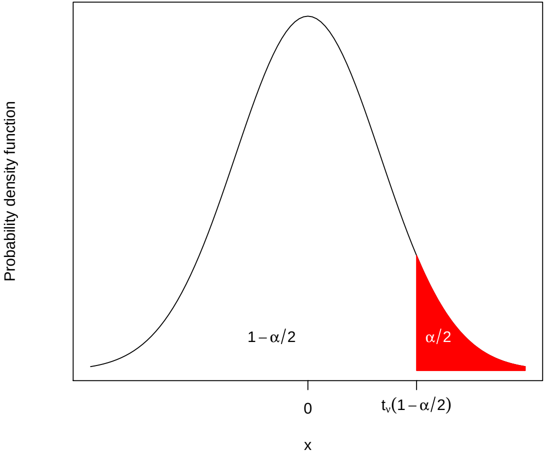
$$L = \text{estimate} - \text{critical value}(1-\alpha) \times \text{standard error}$$

    - One-sided ($H_0 : \mu \geq \mu_0$ vs $H_1 : \mu < \mu_0$): $(-\infty, U)$

$$U = \text{estimate} + \text{critical value}(1-\alpha) \times \text{standard error}$$

Critical values

A related quantity are critical values, e.g. $t_\nu(1-\alpha/2)$.



Let $c = t_\nu(1-\alpha/2)$, then we need $P(t_\nu < c) = 1-\alpha/2$, i.e. the inverse of the cumulative distribution function.

Paired t-test example

In paired t-test example, we had an estimate $\hat{\mu} = 10.5$ and a standard error of 4.3136 with 7 degrees of freedom.

The 95%, i.e. $\alpha = 0.05$, confidence intervals for μ are

- Two-sided ($t_7(.975) = 2.364624$)

$$10.5 \pm 2.364624 \cdot 4.3136 = (0.30, 20.7)$$

- One-sided (positive) ($t_7(.95) = 1.894579$)

$$(10.5 - 1.894579 \cdot 4.3136, \infty) = (2.33, \infty)$$

- One-sided (negative) ($t_7(.95) = 1.894579$)

$$(-\infty, 10.5 + 1.894579 \cdot 4.3136) = (-\infty, 18.7)$$

Two-sample t-test example

In the two-sample t-test example, we had an estimate $\widehat{\mu_1 - \mu_2} = 10.33643$ and a pooled standard error of 0.8190 with 326 degrees of freedom.

The 90%, i.e. $\alpha = 0.10$, confidence intervals for μ are

- Two-sided ($t_{326}(.95) = 1.649541$)

$$10.33643 \pm 1.649541 \cdot 0.8190 = (9.0, 11.7)$$

- One-sided (positive) ($t_{326}(.90) = 1.285149$)

$$(10.33643 - 1.285149 \cdot 0.8190, \infty) = (9.3, \infty)$$

- One-sided (negative) ($t_{326}(.90) = 1.285149$)

$$(-\infty, 10.33643 + 1.285149 \cdot 0.8190) = (-\infty, 11.4)$$

Find critical values using SAS or R

If $\alpha = 0.05$, then $1 - \alpha/2 = 0.975$.

In SAS,

```
PROC IML;
  q = QUANTILE('T', 0.975, 7);
  PRINT q;
  QUIT;
```

In R,

```
q = qt(0.975,7)
```

Both obtain q=2.364.

Equivalence of confidence intervals and pvalues

Theorem

If the $100(1-\alpha)\%$ confidence interval does not contain μ_0, then the associated hypothesis test would reject the null hypothesis at level α, i.e. the pvalue will be less than α.

Examples:

- In the paired t-test example, the one-sided 95% confidence interval for the difference was $(2.33, \infty)$ which does not include 0. Thus the pvalue for the one-sided hypothesis test (with alternative that the difference was greater than zero) was less than 0.05 (it was 0.02) and the null hypothesis was rejected.
- In the two-sample t-test example, the two-sided 95% confidence interval for the difference was $(9.0, 11.7)$ which does not include 0. Thus the pvalue for the two-sided hypothesis test was less than 0.05 (it was < 0.0001) and the null hypothesis was rejected.

Remark Rather than reporting the pvalue, report the confidence interval as it provides the same information and more.

Summary

Two main approaches to statistical inference:

- Statistical hypothesis (hypothesis test)
- Statistical question (confidence interval)