

Lab 14: The t Statistic and Tests of Means

Topics:

- the CLT and means
- the t distribution
- Hypothesis testing and confidence intervals for single population means ( $\mu$ )
- Hypothesis testing and confidence intervals for 2 independent sample population means ( $\mu_1 - \mu_2$ )

=====

We saw the CLT demonstrated for proportions in previous labs. Now, we want to test out if the Central Limit Theorem “works” for means. For our purposes, means will be simply defined as numbers either greater than 1 or less than 0, and having units attached to them (making them quantities). We will do this by randomly sampling from a population of 1000 pennies and noting their ages (in years from minting). Below is a table which has every penny given an ID number (from 000 through 999) and has their ages, as well as the number of pennies with the same age.

Age (yrs)	Count	ID #	Age (yrs)	Count	ID #	Age (yrs)	Count	ID #
0	49	000-048	16	38	643-680	32	6	961-966
1	51	049-099	17	37	681-717	33	6	967-972
2	50	100-149	18	24	718-741	34	1	973
3	85	150-234	19	32	742-773	35	11	974-984
4	47	235-281	20	26	774-799	36	2	985-986
5	61	282-342	21	23	800-822	37	4	987-990
6	29	343-371	22	27	823-849	38	2	991-992
7	29	372-400	23	22	850-871	39	1	993
8	32	401-432	24	19	872-890	40	3	994-996
9	21	433-453	25	10	891-900	46	1	997
10	36	454-489	26	10	901-910	58	1	998
11	38	490-527	27	12	911-922	59	1	999
12	30	528-557	28	13	923-935	<b>TOTAL</b>		<b>1000</b>
13	27	558-584	29	8	936-943			
14	24	585-608	30	12	944-955			
15	34	609-642	31	5	956-960			

First, we must have the data as given above read into R. We will use this by using the `rep(value, number of repeats)` command, as shown below. We need to make 49 zeros, 51 1’s, etc. with their place in the data vector (`data1`) being the ID number. We will use the following R code, located in the website as `pennies.R`:

```
data1 <- c(rep(0,49), rep(1,51), rep(2,50), rep(3,85), rep(4,47), rep(5,61),
          rep(6,29), rep(7,29), rep(8,32), rep(9,21), rep(10,36), rep(11,38),
          rep(12,30), rep(13,27), rep(14,24), rep(15,34), rep(16,38),
          rep(17,37), rep(18,24), rep(19,32), rep(20,26), rep(21,23),
```

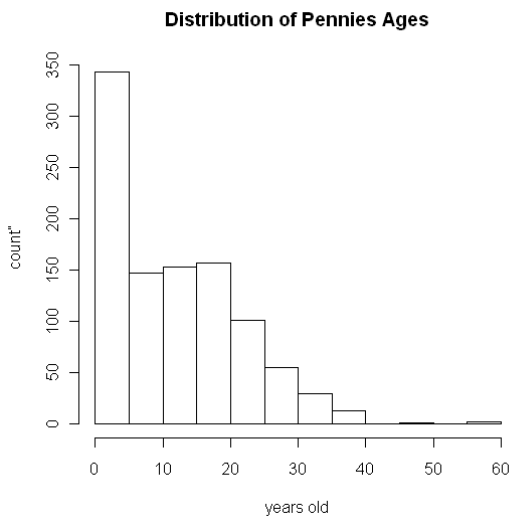
```
rep(22,27), rep(23,22), rep(24,19), rep(25,10), rep(26,10),
rep(27,12), rep(28,13), rep(29,8), rep(30,12), rep(31,5), rep(32,6),
rep(33,6), rep(34,1), rep(35,11), rep(36,2), rep(37,4), rep(38,2),
rep(39,1), rep(40,3), rep(46,1), rep(58,1), rep(59,1))
```

Note that in the actual script I added elements to `data1` slowly, a line at a time, so it was built “correctly”.

Now we want to make “descriptives” (usually, the first task of any statistical investigation), by making summary stats and a histogram, as scripted in R below, along with relevant output.

```
summary(data1)
hist(data1, main="Distribution of Pennies Ages", xlab="years old", ylab="count")
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
 0.00  4.00  11.00 12.26 19.00 59.00
```



We see that this distribution is fairly skewed to the right, with most of the pennies being relatively young.

One of the conditions of the CLT is that we need a “healthy” enough  $n$  on our sampling distribution, in order to have a normal shape. Let’s check that out with our population of pennies’ ages. We will take 100 samples of size 2, 5, 10, 20, and 50 to see what the shape is. The R code to do this might be:

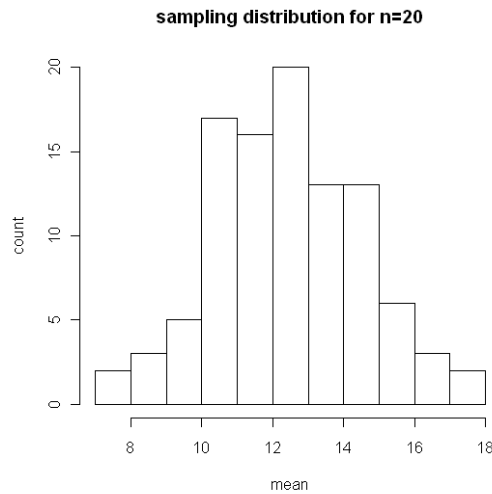
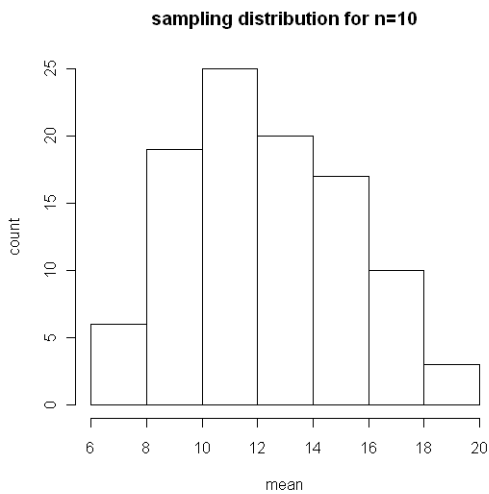
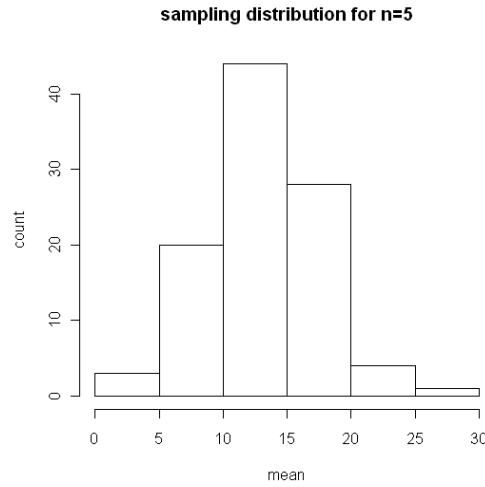
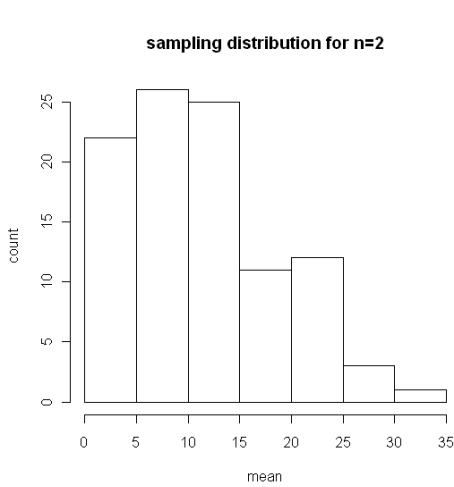
```
samp2 <- c() ; samp5 <- c() ; samp10 <- c() ; samp20 <- c() ; samp50 <- c()
for (i in 1:100) {
  samp <- sample(data1, 2, replace=FALSE)
  samp2[i] <- mean(samp)
}
for (i in 1:100) {
  samp <- sample(data1, 5, replace=FALSE)
  samp5[i] <- mean(samp)
}
for (i in 1:100) {
  samp <- sample(data1, 10, replace=FALSE)
  samp10[i] <- mean(samp)
}
}
```

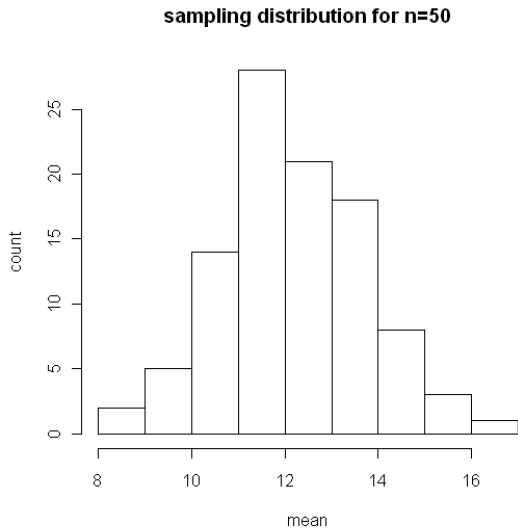
```

for (i in 1:100) {
  samp <- sample(data1, 20, replace=FALSE)
  samp20[i] <- mean(samp)
}
for (i in 1:100) {
  samp <- sample(data1, 50, replace=FALSE)
  samp50[i] <- mean(samp)
}

```

When we make histograms of these various sampling distributions, we see how the shape morphs into a normal as we increase our random samples of the original population. See the below graphs of our sampling distributions.





```

> summary(samp2)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.50   6.50   10.75   11.87   16.62   34.00
> summary(samp5)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.60  10.35   13.00   13.14   15.60   26.80
> summary(samp10)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.80  10.05   12.05   12.25   14.33   18.50
> summary(samp20)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.55  10.92   12.43   12.49   14.00   17.50
> summary(samp50)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.68  11.20   12.04   12.20   13.12   16.32

```

Notice that the mean of the more symmetrical sampling distributions are close to the population mean of 12.26, stated above.

=====

The  $t$  distribution is used for hypothesis testing of means ( $\mu$ ), in the same way as the  $Z$  distribution is used for hypothesis testing of proportions ( $p$ ). The Central Limit Theorem says that the sampling distribution of means has a standard deviation =  $\sigma/\sqrt{n}$ . If we know what  $\sigma$  is we have no problem using the z-statistic in our hypothesis test. The problem is that in actual application we usually don't know what  $\sigma$  is, and we must therefore assume that the sample standard deviation of  $\bar{x}$ ,  $s$ , is close to  $\sigma$ , the population standard deviation. For large sample sizes ( $n$ ) this is probably a good assumption. However, for small sizes of  $n$  we may be far from true in saying that  $s \approx \sigma$ . If our  $s$  is not close to  $\sigma$ , then we are out of line using the z-statistic in our hypothesis test. We need a function which “acts” like z, but has “thicker tails” for smaller and smaller values of  $n$ , so we can cover our error in using  $s$  for the population parameter  $\sigma$ . Along comes the  $t$  distribution to do the task which the  $z$  cannot do.

The  $t$  distribution looks like the bell shaped  $z$  distribution, but for smaller and smaller degrees of freedom (i.e., smaller  $n$ ), the  $t$  has “fatter tails” (and therefore more  $p$ -value

probability for hypothesis tests), so we can cover our discrepancy using  $s$  for  $\sigma$ .

Let us use the values of  $t$  for the 3 standard deviation places (68%, 95%, 99.7%), and see how these quantile  $t$  values get smaller and closer to the center, as the degrees of freedom increase, indicating that the tails are getting smaller and closer to the middle of the  $t$  distribution. See the script and output below.

```
R Console
> # looking at quantiles of t
> #   distribution for various
> #   values of df
> # =====
>
> # first, get 68, 95, and 99 percent values
> #   of t for df=1,4,9,19,49
> #   to correspond to n=2,5,10,20,50
> x <- c(.0015, .025, .17, .57, .975, .9985)
> degoff <- c(1,4,9,19,49)
>
> # now put them in a loop
> for (i in 1:4) {
+ print(qt(x,degoff[i]))
+ }
[1] -212.2050200 -12.7062047 -1.6909077 0.2235265 12.7062047
[6] 212.2050200
[1] -6.4348484 -2.7764451 -1.0823107 0.1880391 2.7764451 6.4348484
[1] -4.0239867 -2.2621572 -1.0075271 0.1814960 2.2621572 4.0239867
[1] -3.4006578 -2.0930241 -0.9787548 0.1787839 2.0930241 3.4006578
~ |
```

Now let us draw the distributions for these same  $t$  distributions, for the same values of  $n=2, 5, 10, 20$  and  $50$ , using the script of problem #1 below.

**[1]** Download and run the script `t_distrib_graph.R`, and notice how the  $t$  distribution begins to look more and more like the  $z$  distribution as the degrees of freedom ( $df$ ) of the  $t$  distribution increase. In fact, as  $df$  approaches infinity, the  $t$  becomes the  $z$ . Copy/paste your resulting graph into WORD. Notice in the script that we used both color distinctions (`col=`) and line format distinctions (`lty=`) for the various graphed distributions.

**[2]** Look at the  $t$ -table in your text book, and estimate at what value of  $df$  you would start to feel comfortable using the  $z$  distribution instead of the  $t$  for hypothesis testing of population means. Write in WORD what your value would be, and a sentence or two saying why you chose what you did.

=====

When doing an hypothesis test for means, a useful R command to use is `t.test(x, mu=..., alt="...")`

In the arguments,  $x$  is the data vector being tested, after the `mu=` is the proposed  $H_0$  mean value, and within the quotes after the `alt=` can be placed either “less”, “greater”, or “two.sided”, depending upon what your alternative hypothesis ( $H_a$ ) is.

A consumer group wishes to see whether the actual mileage of a new SUB matches the

advertised 17 miles per gallon. The group suspects it is lower. To test the claim, the group fills the SUV's tank and records the mileage. This is repeated 11 times. The results are (in mpg): 11.4, 12.0, 13.1, 14.7, 14.7, 15.0, 15.5, 15.6, 15.9, 16.0, 16.8 .

Let us use the R command `t.test` to do a hypothesis test, along with a 95% confidence interval of the true mean for this person's property. The script and results are shown below. Notice that our  $H_a: \mu < 17$  mpg, because we suspect it is lower than the advertised value. This causes the command `alt="less"`. The other options for this command are `alt="greater"` and `alt="two.sided"` .

```
R R Console
> # script to do a single test of means of SUV mpg
> # =====
>
> mpgdata <- c(11.4, 12.0,13.1,14.7,14.7,15.0,15.5,
+             15.6,15.9,16.0,16.8)
> t.test(mpgdata, mu=17, alt="less")

      One Sample t-test

data:  mpgdata
t = -4.5991, df = 10, p-value = 0.0004908
alternative hypothesis: true mean is less than 17
95 percent confidence interval:
 -Inf 15.55133
sample estimates:
mean of x
 14.60909
```

We see that we have very strong evidence to believe that the SUV's claim of 17 mph may be overstated, and that the true average mpg may be less than that.

Now let us do a 95% CI on this same information. The input and output are shown below.

```
R R Console
> t.test(mpgdata, conf.level=.95)

      One Sample t-test

data:  mpgdata
t = 28.1014, df = 10, p-value = 7.564e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 13.45075 15.76743
sample estimates:
mean of x
 14.60909
```

We see that we are 95% confident that the true mpg of this SUV, according to our information, is between about 13.45 and 15.77 mpg. We need to realize, with our results, that our conclusions may be a bit biased or incorrect, because we have a rather small  $n$  to do this test ( $n=11$ ) and may have randomness problems associated with our sample values

obtained in those 11 tests.

[3] The web site has a data set called `normtemp.txt`, which is a listing of a group of randomly picked subjects' body temperatures (`temperature`) and heart rates (`hr`). A third variable, their gender (`gender`), is also listed. The researcher wants to challenge the belief that normal body temperature is 98.6°F, thinking that the true mean temperature of humans is less than that. Using R, do an hypothesis test of this problem, and state a conclusion. Also, remark on the appropriateness of the assumptions needed for this test (i.e., randomness,  $n$  size, etc.).

[4] The researcher knows that some scientists have suggested that 98.2°F is a better value for the mean body temperature. Perform a 95% confidence interval, interpret your interval, and write a conclusion as to feasibility of the new value. Again, remark on appropriateness of the test assumptions, using the researcher's data.

=====

Using this same temperature data, we want to see if there is a difference in body temperature between men and women. We can use R for this test of differences in independent means samples. We first read in the data as `data1`, where the first column is the `temperature` column, and second column is `gender`. Then we separate the temperatures for men and women with the command:

```
males <- data1[,1][data1[,2]=="Male"] and  
females <- data1[,1][data1[,2]=="female"].
```

Note that this command makes `males` only those temperatures from column 1 which have "Male" in column 2, and likewise for `females`.

Be careful with spelling and case sensitivity on these commands, to avoid error messages.

We first want to test the hypothesis that women have a higher temperature than men, on average, and next do a 95% confidence interval on the difference between temperatures between men and women, using this data. The hypothesis test command we use to test

$\mu_{males} - \mu_{females}$  is

```
t.test(males, females, mu=0, alt="less", var.equal=TRUE)
```

That last entry (`var.equal=TRUE`) is stated because we are seeing that the variances in both samples are relatively equal. We could do a `summary(males)` and `summary(females)`, or plot histograms, etc., to see that. Note that the default value for `var.equal=` is `FALSE`. Our input and output is shown below.

## R Console

```
> data1 <- read.table("normtemp.txt", header=TRUE)
> males <- data1[,1][data1[,2]=="Male"]
> females <- data1[,1][data1[,2]=="female"]
> t.test(males, females, mu=0, alt="less", var.equal=TRUE)

      Two Sample t-test

data:  males and females
t = -2.2854, df = 128, p-value = 0.01197
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -0.07955046
sample estimates:
mean of x mean of y
 98.10462  98.39385
```

Our results of the hypothesis test show that we have strong evidence that females have a different, higher average temperature than males, assuming our samples are SRS and independent, or at least representative of our population.

Below is our confidence interval, representing  $\bar{y}_{females} - \bar{y}_{males}$  using the command `t.test(females, males, var.equal=TRUE, conf.level=0.95)`

## R Console

```
> t.test(males, females, var.equal=TRUE, conf.level=.95)

      Two Sample t-test

data:  males and females
t = -2.2854, df = 128, p-value = 0.02393
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.53963938 -0.03882216
sample estimates:
mean of x mean of y
 98.10462  98.39385
```

We see that our confidence level does not include 0, and that females have a warmer body temperature than males.

**[5]** Perform, in WORD, a 95% confidence interval on the difference between heart rates of men and women, using our data, and also perform a hypothesis test that womens' heart rates are faster than mens'. Be sure to check that the conditions for confidence intervals and hypothesis tests for 2 sample means is met, and that the variance is about the same, by doing summaries and/or graphs of the data.