Something you need to know

Synferlo

May. 4, 2021

1 Sampling Distribution

1.1 Std. deviation vs. Std. error

Two important parameters in each distribution: mean, standard deviation.

Given the sample data, we want to estimate the distribution of the population, i.e., want to know which distribution generate the data.

By estimating the distribution, we estimate the mean and the std deviation. The estimated std deviation of a sampling distribution is referred to as a std error.

$$\text{Std. deviation} = \sqrt{\text{variance}}$$

1.2 Estimate sampling distribution

Mean of the sampling distribution should be the same as the true mean, $\mathbb{E}(\mu_X) = \mu$. And the std. error, s_X:
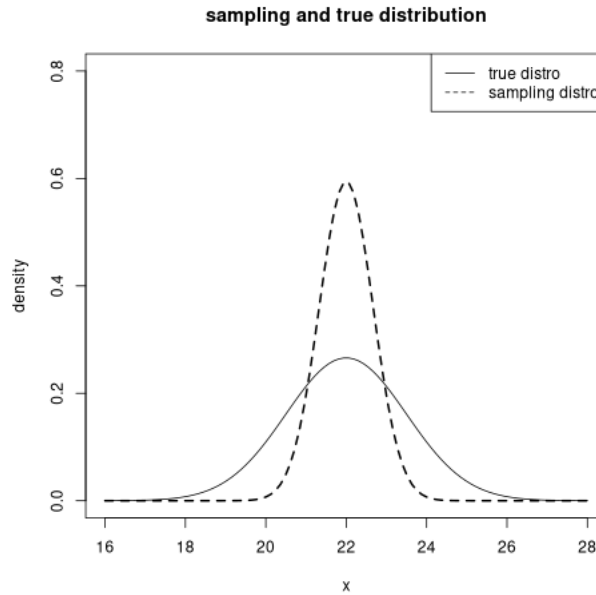
$$s_X = \sqrt{\frac{\text{variance}}{n}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Suppose we generate a sample, with size 5, from $Normal(\mu = 22, \sigma^2 = 1.5^2)$, the standard error for the sampling distribution would be

$$s_X = \frac{\sigma}{\sqrt{n}} = \frac{1.5}{\sqrt{5}} \approx 0.671$$

Then we can use R to plot the sampling and true distribution.

Figure 1: Sampling vs. True distribution

Clearly, the sampling distribution is taller and skinner, while the true distribution is flatter.

1.2.1 Confident Interval for the mean

By doing this, must assume the data is normally distributed.

Then,

Step 1: Compute sample mean, μ_X and std deviation, s_X. (page 394 in Tilman's R code book)

Step 2: Compute sample std. error.

$$se = \frac{s_X}{\sqrt{n}}$$

Step 3: Since we assume obs are normally distributed and we are using s_X rather σ_X, t-distribution with $n-1$ degree of freedom is the appropriate method, where n is the sample size.

Step 4: Choose the confident interval, e.g., 95% . It says $\alpha = 0.05$. Hence the critical value

is

$$\text{cumulative probability } = 1 - \frac{\alpha}{2} = 0.975$$

Now we have the cumulative probability, p, for the qt() function in R. We can use the following code to compute the critical value (or called the quantile value in R).

```
α = 0.975

#You need to substitute n by the actual sample size.

critical_value = qt(p = α, df = n - 1)

print(critical_value)

#it is the value on the horizontal axis, if you plot the density of t-distribution.
```

So the area within the confidence interval(CI) is $prob. = .95$
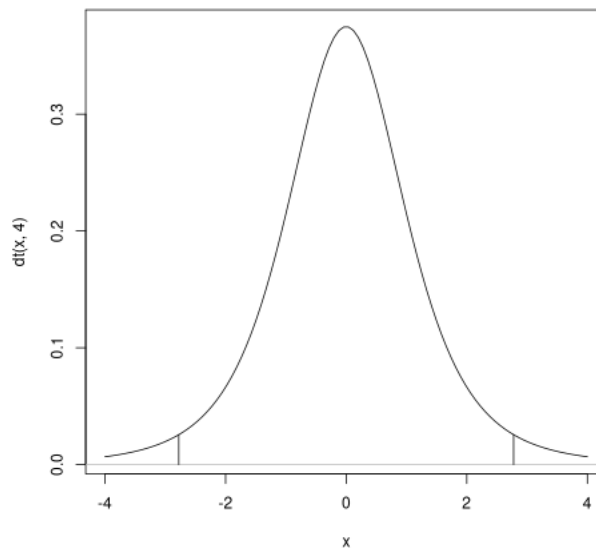
Step 5: Now you have sample mean (μ_X), std error (se), and the critical value for .95 confidence interval. Then you can make the following inference:

I am in 95% confidence saying that the true mean of the population lies somewhere between

$$\mu_X - se \times \text{critical\_value} < \mu < \mu_X + se \times \text{critical\_value}$$

2 Hypothesis Test

2.1 null and alternative hypothesis

2.1.1 Null Hypo.

Null Hypothesis: H_0. The claim that is assumed to be true.

2.1.2 Alt. Hypo.

All. hypo.: H_1. The conjecture that you are testing for, against the null.

Three types of tests based on different H_1:

1. lower-tailed test: When H_1 is defined in terms of a less-than statement, with $<$. It is one-sided.

2. upper-tailed test: When H_1 is defined in terms of a greater-than statement, with $>$. It is one-sided.

3. two-tailed test: When H_1 is defined in terms of a different-to statement, with $\neq$. It is two-sided.

2.2 Test statistics

The statistic that is compared to the appropriate standardized sampling distribution to yield the p-value.

2.3 p-value

The meaning for p-value in different types of test:

1. In a lower-tailed test: p-value is a left-hand tail probability from the sampling distribution of interest.

2. In an upper-tailed test: a right-hand tail prob.

3. In two-sided test, the sum of left and right tail probability. This is equivalent to two times the area in one tail when the sampling distribution is <u>symmetric</u>.

Consider and example:
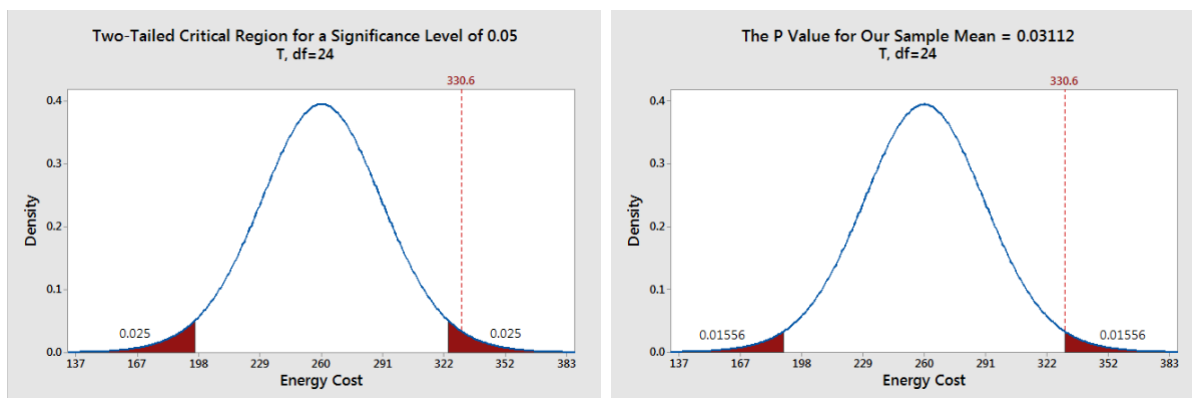
H_0: the population mean is 260

H_1: the population mean differs from 260

After calculation, we get the sample mean is 330.6

Red area in the left figure stands for the area of our 95% CI. Clearly, our sample mean is outside the CI.

Red area in the right figure stands for the p-value for this sample mean. Note, we can compute this by using pnorm() in R. The p-value says the prob. of obtaining a <u>sample mean</u> that is as extreme (or even far away from the true mean, 260) as what we got (330.6) is $0.01556 \times 2 = 0.03112$.

Clearly, when the sample mean appears outside the 95% CI, the p-value is also less than the significance level, $0.03112 < 0.05$. Or we say p-value is less than the significance level.

Since our sample mean is outside the 95 CI, we CANNOT say that the population mean is 260 with 95% confidence. That's why we <u>reject the null when p-value $<$ the significance level.</u>

Remember, p-value is calculated based on the position of your sample mean. If the p-value is less than the significance level, <u>your sample mean is outside the CI</u>. Hence, we reject the null.

2.4 Comments on Hypothesis Test

1. p-value never provides "proof" of either H_0 or H_1 being truly correct.

Rejecting H_0 merely implies that the sample data suggest H_1 ought to be preferred.

2.5 Testing Means (single mean)

How to test the mean with t-test? Let's do this.

Suppose you have a sample from the population with mean = 80. You want to test if the true mean, i.e., population mean, is 80 with 95 confidence.

$$H_0 : \mu = 80$$

$$H_1 : \mu \leq 80$$

By doing this, we need to:

step 1: compute sample mean$(\overline{X})$ and the sample std deviation, or SD in short (s).

mean = 80.10174, SD = s = 1.651756.

step 2: compute test statistic T:

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \approx 0.3895706$$

The dashed line stands for t = 0.3895706. Note, the sample SD, $s/\sqrt{n}$, is the estimated standard error of the mean.

T follows a t-distribution with $n - 1$ degree of freedom.

step 3: Note, H_1 suggests that this is a left-tailed test. So, the p-value is provided as the area

under the sampling distribution (a t-distribution with DOF = 39) to the left of a vertical line at T.

Now, compute the p-value $= 0.6505133$. It says the area at the left side of the dashed line is about 0.65 (The cumulative prob.).

step 4: Compare the p-value with significance level, α. Clearly, p-value is much greater than 0.05. We fail to reject the null. It pretty much makes sense because the sample we use is generated from a normal distribution with mean $= 80$. It is very unlikely we can reject the null. The code is below.

Test the mean of 80
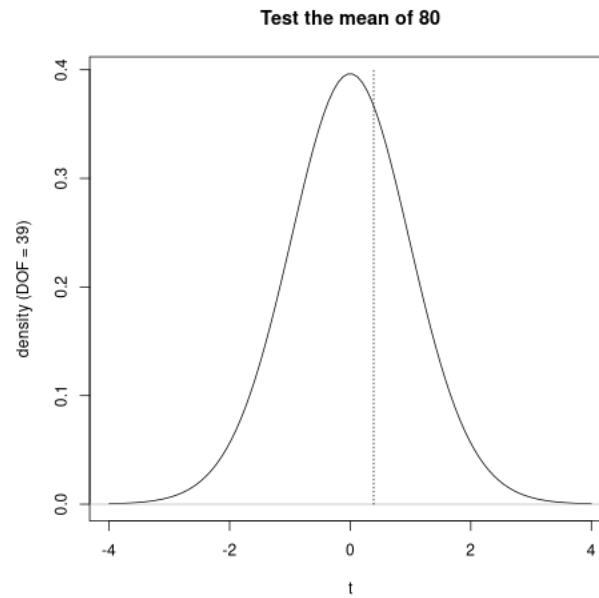

```
# generate a sample
n = 40
# make sure we have the same generated numbers, set the seed
set.seed(5)
sample = rnorm(n, 80, 1.5)
# compute the sample mean and SD
sample_mean = mean(sample)                          # X = [1] 80.10174
# sample SD: s
```

8

```r
sample_sd = sd(sample)                              # s = [1] 1.651756
sample_se = sample_sd/sqrt(n)    # [1] 0.2611655


# T statistics:
sample_T = (sample_mean - 80)/sample_se             # [1] 0.3895706


# compute the p-value, i.e., the cumulative prob. using pt()
# remember, sample_T is t value on the horizontal axis of the density.
sample_p_value = pt(sample_T, n-1)                  # [1] 0.6505133


# Clearly, p-value is much greater than 0.05. We fail to reject
# the null.



# Method 2
# R has a built in function for t-test:
# t.test(x = sample, mu = mean, alternative = '')
# alternative can be two.sided, less, greater



t.test(x = sample, mu = 80, alternative = 'less')


#                 One Sample t-test
#
#       data:  sample
#       t = 0.38957, df = 39, p-value = 0.6505
#       alternative hypothesis: true mean is less than 80
#       95 percent confidence interval:
#               -Inf 80.54177
```

```
#      sample estimates:
#      mean of x
#       80.10174
```

2.6   Testing Means (two mean)

Sometime, testing for one mean is not enough. You may want to compare the means of two distinct groups of measurements, which boils down to a hypothesis test for the true difference between two means, μ_1, μ_2.

Two groups of data relate to each other affects the specific form of standard error for the difference between two sample means and therefore the test statistics itself. Normally, H_0 would be $\mu_1 = \mu_2$.

Case 1:

When you cannot assume the <u>variances</u> of the two populations are <u>equal</u>, then you perform the <u>unpooled</u> version of the two-sample t-test.

Case 2:

If you can safely assume <u>equal variances</u>, then you can perform a <u>pooled</u> two-sample t-test, which <u>improve the precision of the result</u>.

2.6.1   Unpooled two-sample t-test

Suppose we have two samples from $N(76, 2)$, $N(80, 2)$, with size $n_1 = 44$, $n_2 = 31$. Clearly, we manually make the population of these two samples with $\mu_1 = 76$, $\mu_2 = 80$. Now let's use t-test to verify the following null hypothesis. (Note, we are testing the mean of the

population.)

$$H_0 : \mu_1 - \mu_2 = 0$$

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

We can use Welch's t-test:

Form the t ststistics for two-sample case,

$$T = \frac{\mu_{X_1} - \mu_{X_2} - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where μ_0 is the null value of the interest, which is $\mu_0 = \mu_1 - \mu_2 = 0$.

And this t statistic is approximately following a t-distribution with df degree of freedom, where

$$df = \left[\frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1-1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2-1}} \right]$$

The above equation for df is called the Welch-Satterthwaite equation.

Note, the numerator of T should be consistent with the hypothesis. Here, we are testing for either $\mu_1 - \mu_2 = 0$ or $\mu_1 - \mu_2 < 0$, so in T we write $\mu_{X_1} - \mu_{X_2}$. It is a lower-tailed test

If you write your hypothesis in this way,

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 > 0.$$

Then it will be a upper-tailed test. And you want to check to right tail, rather the left one .

In method 1, I compute the t statistics and p-value by myself. $T = -299.4002$, p-value $= 2.560795e{-107}$. It is clearly less than 99% confidence interval. Hence, we can reject the null. It says that the population means are not likely to be equal based on our sample.

You can check the figure below. At left tail, the probability is almost zero even when t is

around 4. Hence, out $t = -299$ must be closed to 0. And this is verified by the p-value.
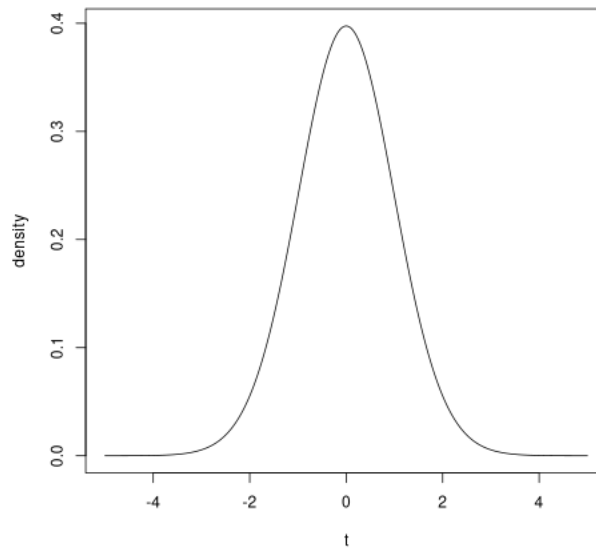
Instead of writing the code by myself, R also have a built-in function for Welch's t-test.

```
# Method 1
# Generate data from N(76,2), N(80,2)
set.seed(5)
n1 = 44
n2 = 31
sample1 = rnorm(n1, 76, 2)
sample2 = rnorm(n2, 80, 2)



# Compute sample mean and se, i.e., s
sample1.mean = mean(sample1)          # [1] 76.24008
sample2.mean = mean(sample2)          # [1] 79.66198
```

```r
sample1.s = sd(sample1)/n1                            # [1] 0.05069484
sample2.s = sd(sample2)/n2                            # [1] 0.04731536



# Compute df:


df.upper = ((sample1.s**2/n1) + (sample2.s**2/n2)) ** 2
df.lower.left = (sample1.s**2/n1)**2/(n1-1)
df.lower.right = (sample2.s**2/n2)**2/(n2-1)
df.lower = df.lower.left + df.lower.right


df = df.upper/df.lower                                # [1] 67.39422



# Compute t statistics
t.upper = sample1.mean - sample2.mean
t.lower = sqrt(
            (sample1.s**2/n1) + (sample2.s**2/n2)
)
# It is negative because we are doing a left-tailed test.
sample.t = t.upper/t.lower              # [1] -299.4002


# p-value
pt(sample.t, df)                        # [1] 2.560795e-107


# Suppose we use 99 Confidence Interval, p-value here is clearly
# less than 0.01
# Great, we reject the null. Hence miu_1 < miu_2. It is being verified.
```

```
png('figures/t-test_two_mean.png')
x = seq(-5,5, length = 2000)
plot(x,
            dt(x, df),
            type = 'l',
            xlab = 't',
            ylab = 'density'
)
dev.off()



# Method 2
# Built-in Welch's t-test
t.test(x = sample1, y = sample2, alternative = 'less', conf.level = 0.99)




#             Welch Two Sample t-test
#
#     data:  sample1 and sample2
#     t = -8.0105, df = 72.718, p-value = 6.891e-12
#     alternative hypothesis: true difference in means is less than 0
#     99 percent confidence interval:
#         -Inf -2.405759
#     sample estimates:
#     mean of x mean of y
#      76.24008  79.66198
```