Statistical Tests (1)

MAE301 Applied Experimental Statistics

Yi Ren, Yabin Liao

School for Engineering of Matter, Transport Energy
Arizona State University

September 15, 2015

Outline

review: sample mean, sample variance

A **statistic** is a function of the random variables in a random sample.

Sample mean (average): $\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$

Sample variance : $s^2 := \frac{1}{\mathbf{m-1}} \sum_{i=1}^{m} (x_i - \bar{x})^2$

Sample standard deviation : $s := \sqrt{\frac{1}{\mathbf{m-1}} \sum_{i=1}^{m} (x_i - \bar{x})^2}$

exercise

To estimate the average height (at the shoulders) of dogs, Joe measured the heights of his five dogs and the results are 600 mm, 470 mm, 170 mm, 430 mm and 300 mm, respectively. Determine the sample mean and sample standard deviation of the measurements. [394 mm, 164 mm]

review: central limit theorem

Central limit theorem: If $\bar{X}$ is the mean of a random variable of size n taken from a population with mean μ and variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{1}$$

as $n \to \infty$, is the standard normal distribution $N(0, 1)$.

The sample size $n = 30$ is a guideline to use for the central limit theorem. The normal approximation will generally be good if $n \geq 30$. If $n < 30$, the approximation is good only if the population is not too different from a normal distribution.

exercise

The life of a type of electric light bulb can be modeled with exponential probability function

$$f(x) = \begin{cases} 0.001e^{-0.001x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where x is the life in hours.

1. Determine the mean and variance of the light bulb life
2. If 40 light bulbs are installed in a building, what is the probability that the average life of the bulbs is less than 1200 hours?

statistical inference on population mean

What can we say about the population (true) mean (μ) when we see the sample mean ($\bar{x}$)? Or what is the reasonable guess for the population mean?

Recall that the expectation of $\bar{x}$ is μ, and its variance is σ^2/n. By central limit theorem, $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ is approximately standard normal when $n > 30$. We can find the range of μ **when** σ^2 **is known**.

exercise

An important manufacturing process produces cylindrical component parts for the automotive industry. It is important that the process produces parts having a mean of 5.0 mm. The engineer involved speculates that the population mean is 5.0 mm.

An experiment is conducted in which 100 parts produced by the process are selected randomly and the diameter measured on each. It is known that the population standard deviation $\sigma = 0.1$ mm. The experiment indicates a sample average diameter $\bar{x} = 5.027$mm.

Does this sample information appear to support or refute the engineers speculation?

statistical inference on population mean

The previous example is inference on a single population mean. A far more important application involves two populations.

For example, two manufacturing methods, 1 and 2, are compared. The basis for that comparison is $\mu_1 - \mu_2$, the difference in the population means.

Let the statistic $\bar{X}_1$ represent the mean of a random sample of size n_1 selected from a population with mean μ_1 and variance σ_1^2, and the statistic $\bar{X}_2$ represent the mean of a random sample of size n_2 selected from a population with mean μ_2 and variance σ_2^2.

If the populations are of normal distribution, or the sample size n is large (i.e., central limit theorem)

$$\bar{X}_{1,2} = N\left(\mu_{1,2}, \frac{\sigma_{1,2}^2}{n_{1,2}}\right) \tag{3}$$

statistical inference on population mean

If independent samples of large size n_1 and n_2 are draw at random from two population with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2, respectively, then the sample distribution of the difference of means, $\bar{X}_1 - \bar{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2, \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{4}$$

Hence

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{5}$$

is approximately a standard normal distribution variable.

Note that if both populations are normal, then $\bar{X}_1 - \bar{X}_2$ has a normal distribution regardless of the sample size.

exercise

Two independent experiments are being run in which two different types of paints are compared. Eighteen specimens are painted using Type A and the drying time, in hours, is recorded on each. The same is done with Type B. The population standard deviations are both known to be 1.0.

Assuming that the mean drying time is equal for the two types of paint, and $\bar{X}_A$ and $\bar{X}_B$ are average drying times, find $P(\bar{X}_A - \bar{X}_B > 1.0)$ and $P(\bar{X}_A - \bar{X}_B > 0.25)$.

t-distribution

Recall that the previous inference on population mean requires σ to be known. However, it is rarely the case that we actually know σ. What do we do then?

A natural way would be approximating the population standard deviation σ with sample standard deviation s.

The random variable

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \tag{6}$$

follows the Student's t-distribution with $\nu = n - 1$ degree of freedom.

Note that the T statistic is not a result of the central limit theorem and samples must come from a normal distribution in order for T to follow a t-distribution.

t-distribution curve

- Similar to the standard normal distribution Z, the t-distribution T is also bell shaped and symmetric.
- The t-distribution is more variable. The variance of T depends on the sample size n and is always greater than 1.
- When $n \to \infty$, the two distributions become the same.

By convention, we let t_α to represent the t-value above which the area equals to α (upper-tail probability):

$$P(T > t_\alpha) = \alpha \tag{7}$$

Due to symmetry, $t_{1-\alpha} = -t_\alpha$.

exercise

Calculate t values with 11 degree of freedom (i.e., a random sample of size 12 selected from a normal distribution)

1. Determine $t_{0.01}$, $t_{0.975}$, $P(-t_{0.025} < T < t_{0.05})$
2. Find k such that $P(k < T < -1.796) = 0.045$

exercise

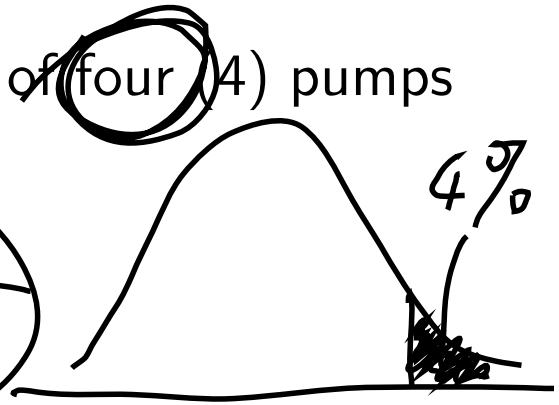
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1372.2 - 1350}{114.7/\sqrt{6}}$$

The manufacturer of general aircraft vacuum pumps wishes to study the failure time of its product. It is known that the mean failure time is 1350 hours.

To obtain the variance information, six pumps are tested to failure with these results (in hours of operation): 1272, 1384, 1543, 1465, 1250, and 1319. ($\bar{x} = 1372.2$, $s = 114.7$)

Find the probability that the average failure time of four (4) pumps is greater than 1500 hours.

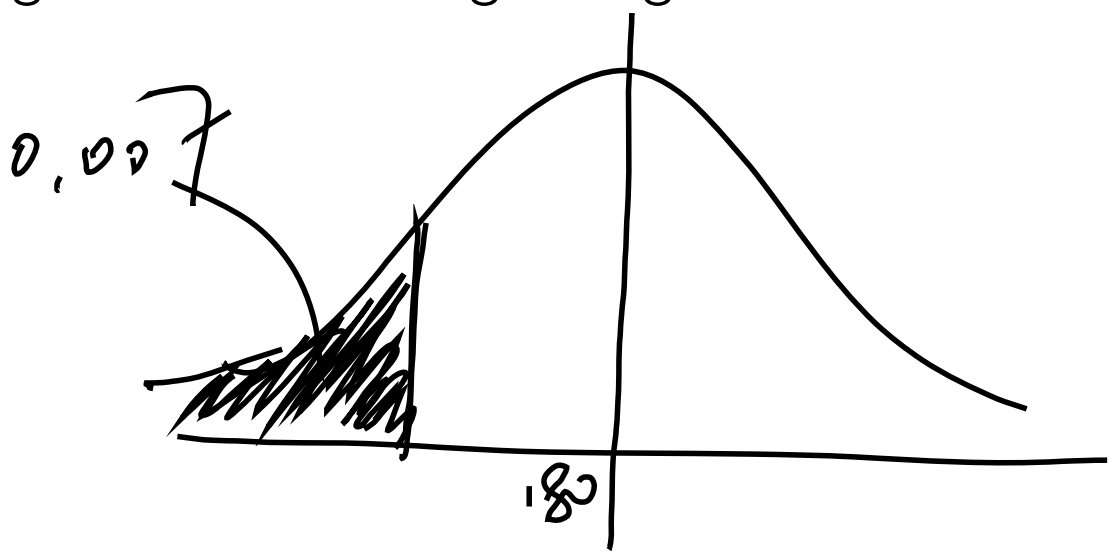
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1500 - 1350}{114.7/\sqrt{4}}$$



exercise

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{169.5 - 180}{5.7/\sqrt{5}}$$

A certain kind of steel cables has a mean breaking strength of 180 lbs. If five pieces of the steel cable (randomly selected from different rolls) have a mean breaking strength of 169.5 lbs with a sample standard deviation of 5.7 lb, what conclusions can you draw regarding the mean braking strength of the cable?



two-sample t-test

Recall that if independent samples of large size n_1 and n_2 are draw at random from two population with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2, respectively, then the sample distribution of the difference of means, $\bar{X}_1 - \bar{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2, \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{8}$$

Hence

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{9}$$

is approximately a standard normal distribution variable.

Note that if both populations are normal, then $\bar{X}_1 - \bar{X}_2$ has a normal distribution regardless of the sample size.

two-sample t-test (cont.)

When σ_1^2 and σ_2^2 are unknown but equal, we can use the following approximation based on sample variances s_1^2 and s_2^2:

$$\sigma^2 \approx s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \tag{10}$$

where s_p^2 is the pooled sample variance, which is essentially a weighted average of the two sample variances with the weights being the degrees of freedom.

two-sample t-test (cont.)

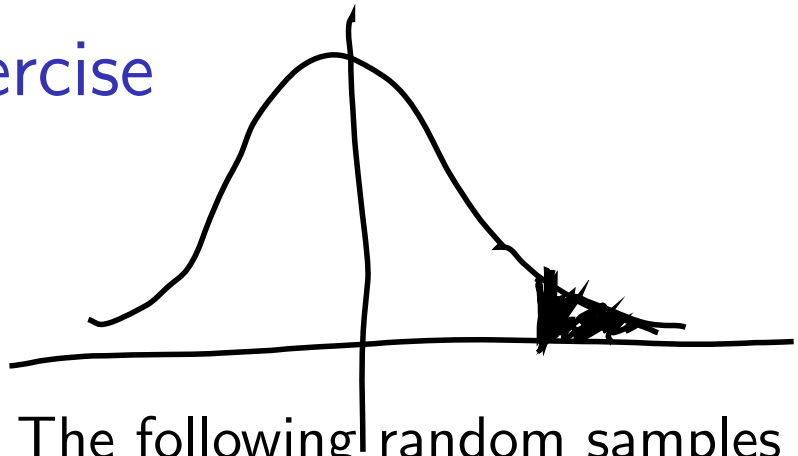
The random variable:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2(1/n_1 + 1/n_2)}} \tag{11}$$

follows the t-distribution with degree of freedom $\nu = n_1 + n_2 - 2$.

We can use T to determine whether we can accept the hypothesis that the two samples have the same mean.

exercise



The following random samples are measurements of the heat-producing capacity (in millions of calories per ton) of specimens of coal from two mines:

Mine 1	8260	8130	8350	8070	8340	
Mine 2	7950	7890	7900	8140	7920	7840

Perform a test to determine if the difference between the heat producing capacities is significant.

two-sample t-test (cont.)

What if σ_1^2 and σ_2^2 are unknown and unequal? In this case the statistic T' will be

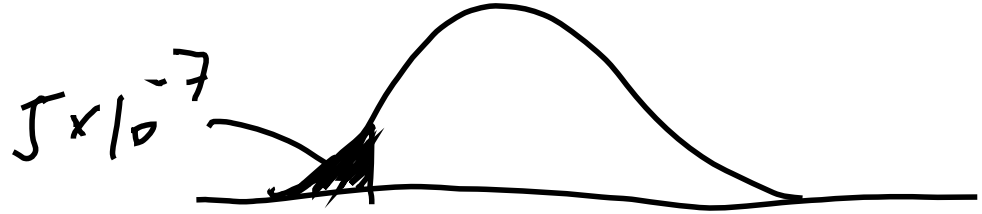
$$T' = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2/n_1 + s_2^2/n_2)}}, \tag{12}$$

which has approximately a t-distribution with ν degree of freedom, where

$$\nu = \left\lfloor \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)} \right\rfloor. \tag{13}$$

Note that $\lfloor x \rfloor$ is the largest integer smaller than x.

exercise



If from two samples (each with 9 measurements): $\bar{x}_1 = 1.334$, $s_1^2 = 0.4548$, $\bar{x}_2 = 3.766$, $s_2^2 = 0.1089$.

Note that $s_1^2 > 4s_2^2$. We should not assume the population variances are equal. Thus the degree of freedom is:

$$\nu = \left\lfloor \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)} \right\rfloor = \lfloor 11.62 \rfloor = 11. \tag{14}$$

T' statistic is

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2/n_1 + s_2^2/n_2)}} = -9.71 \tag{15}$$

Then check if this t' value falls within the reasonable range based on the assumptions made.

Mann-Whitney *u*-test

But we don't know if the two samples are from normal distributions... In this case we will calculate the U statistic (what is this?)

u-test examines whether one distribution is stochastically larger than the other (one-tail), or whether the two distributions are the same (two-tail).

u-test properties

1. When the two distributions are not normal
2. When data is ordinal (e.g., rankings), U-test is the logical choice.
3. More robust to outliers than t-test (since it only uses ranking)

summary of the class

- Z statistic (when true variance is known, distribution is normal or sample size is large) and T statistic (when true variance is unknown and distribution is normal)
- one sample t-test and two sample t-test (equal and unequal variances)
- u-test (when distributions are not normal or samples are ordinal)

Python code for demos in the class

```
# t distribution vs standard normal
import numpy as np
from scipy.stats import t
from scipy.stats import norm
import matplotlib.pyplot as plt
fig, ax = plt.subplots(1, 1)
df = 100 # DOF of t
x = np.linspace(t.ppf(0.001, df),t.ppf(0.999, df), 100)
ax.plot(x, t.pdf(x, df),'r-', lw=5, alpha=0.8, label='t pdf')
ax.plot(x, norm.pdf(x),'k-', lw=5, alpha=0.3, label='norm pdf')

# exercise 1 on t distribution
df = 12-1 # 12 samples
import numpy as np
from scipy.stats import t
t.ppf(0.01,df)
t.ppf(0.975,df)
t.ppf(t.cdf(-1.796,df)-0.045,df)

# exercise 2 on t distribution
import numpy as np
from scipy.stats import t
sample = [1272, 1384, 1543, 1465, 1250, 1319]
xbar = np.mean(sample)
s = np.std(sample,ddof=1)
# first check if the sample is a rare event (if so, we will need to question the true mean)
df1 = 6-1 # 6 samples
t.cdf((xbar-1350)/(s/np.sqrt(6)),df1) # ok good
# then calculate the probability for average of 4 pumps to be over 1500
df2 = 4-1 # 4 samples
1-t.cdf((1500-1350)/(s/np.sqrt(4)),df2)
```

Python code for demos in the class

```
# exercise 3 on t distribution
import numpy as np
from scipy.stats import t
xbar = 169.5
s = 5.7
# first check if the sample is a rare event (if so, we will need to question the true mean)
df = 5-1 # 5 samples
t.cdf((xbar-180)/(s/np.sqrt(5)),df)
#cdf = 0.007 < 0.05, not likely to happen, the original mean could be wrong

# exercise on two sample t test
import numpy as np
from scipy.stats import t
sample1 = [8260,8130,8350,8070,8340]
sample2 = [7950,7890,7900,8140,7920,7840]
xbar1 = np.mean(sample1)
s1 = np.std(sample1,ddof=1)
df1 = len(sample1)-1
xbar2 = np.mean(sample2)
s2 = np.std(sample2,ddof=1)
df2 = len(sample2)-1
sp2 = (((df1-1)*s1**2)+((df2-1)*s2**2))/(df1+df2-2)
z = (xbar1-xbar2)/(np.sqrt(sp2*(1.0/df1+1.0/df2)))
t.cdf(z,df1+df2-2)
# p-value = 2*(1 -t.cdf(z,df1+df2-2)) = 0.002

# use two sample t test directly to validate the above result
from scipy.stats import ttest_ind
ttest_ind(sample1,sample2, equal_var=True)
```