Statistical Methods and Data Analysis I
Lecture 15: Student's T-Test

Oleg Goldshmidt
oleg.goldshmidt@post.idc.ac.il

Arison School of Business
Interdisciplinary Center (IDC)
Herzliya, Israel

May 12, 2019

Do Two Distributions Have the Same Mean?

Do Two Distributions Have the Same Mean?

- **A very common question**:
  - one sample may be gathered before some *event*, the other sample — after
    - "event" may be treatment, change of a control parameter, etc.
    - we want to know whether or not it made a difference...

Do Two Distributions Have the Same Mean?

- **A very common question**:
  - one sample may be gathered before some *event*, the other sample — after
    - "event" may be treatment, change of a control parameter, etc.
    - we want to know whether or not it made a difference...
- **First thought**: check how many *standard deviations* one sample mean is from the other.

Do Two Distributions Have the Same Mean?

- **A very common question**:
  - one sample may be gathered before some *event*, the other sample — after
    - "event" may be treatment, change of a control parameter, etc.
    - we want to know whether or not it made a difference...
- **First thought**: check how many *standard deviations* one sample mean is from the other.
  - may be a useful thing to know: related to the *importance* or *strength* of the difference in means *if that difference is significant*.

Do Two Distributions Have the Same Mean?

- **A very common question**:
  - one sample may be gathered before some *event*, the other sample — after
    - "event" may be treatment, change of a control parameter, etc.
    - we want to know whether or not it made a difference...
- **First thought**: check how many *standard deviations* one sample mean is from the other.
  - may be a useful thing to know: related to the *importance* or *strength* of the difference in means *if that difference is significant*.
  - says *nothing* about whether the difference is statistically significant.

Do Two Distributions Have the Same Mean?

- **A very common question**:
  - one sample may be gathered before some *event*, the other sample — after
    - "event" may be treatment, change of a control parameter, etc.
    - we want to know whether or not it made a difference...
- **First thought**: check how many *standard deviations* one sample mean is from the other.
  - may be a useful thing to know: related to the *importance* or *strength* of the difference in means *if that difference is significant*.
  - says *nothing* about whether the difference is statistically significant.
    - The difference in means may be very small compared to the standard deviation, and yet very significant if the sample is large, or it may be quite large but not significant if the data are sparse.

Do Two Distributions Have the Same Mean?

- **A very common question**:
  - one sample may be gathered before some *event*, the other sample — after
    - "event" may be treatment, change of a control parameter, etc.
    - we want to know whether or not it made a difference...
- **First thought**: check how many *standard deviations* one sample mean is from the other.
  - may be a useful thing to know: related to the *importance* or *strength* of the difference in means *if that difference is significant*.
  - says *nothing* about whether the difference is statistically significant.
    - The difference in means may be very small compared to the standard deviation, and yet very significant if the sample is large, or it may be quite large but not significant if the data are sparse.
  - **"strength"** and **"significance"** are *distinct*!

Do Two Distributions Have the Same Mean?

- **A very common question**:
  - one sample may be gathered before some *event*, the other sample — after
    - "event" may be treatment, change of a control parameter, etc.
    - we want to know whether or not it made a difference...
- **First thought**: check how many *standard deviations* one sample mean is from the other.
  - may be a useful thing to know: related to the *importance* or *strength* of the difference in means *if that difference is significant*.
  - says *nothing* about whether the difference is statistically significant.
    - The difference in means may be very small compared to the standard deviation, and yet very significant if the sample is large, or it may be quite large but not significant if the data are sparse.
  - **"strength"** and **"significance"** are *distinct*!
- **What determines statistical significance?**

Do Two Distributions Have the Same Mean?

- **A very common question**:
  - one sample may be gathered before some *event*, the other sample — after
    - "event" may be treatment, change of a control parameter, etc.
    - we want to know whether or not it made a difference...
- **First thought**: check how many *standard deviations* one sample mean is from the other.
  - may be a useful thing to know: related to the *importance* or *strength* of the difference in means *if that difference is significant*.
  - says *nothing* about whether the difference is statistically significant.
    - The difference in means may be very small compared to the standard deviation, and yet very significant if the sample is large, or it may be quite large but not significant if the data are sparse.
  - **"strength"** and **"significance"** are *distinct*!
- **What determines statistical significance?**
  - Standard error! It takes into account the sample size!

Do Two Distributions Have the Same Mean?

- **A very common question**:
  - one sample may be gathered before some *event*, the other sample — after
    - "event" may be treatment, change of a control parameter, etc.
    - we want to know whether or not it made a difference...
- **First thought**: check how many *standard deviations* one sample mean is from the other.
  - may be a useful thing to know: related to the *importance* or *strength* of the difference in means *if that difference is significant*.
  - says *nothing* about whether the difference is statistically significant.
    - The difference in means may be very small compared to the standard deviation, and yet very significant if the sample is large, or it may be quite large but not significant if the data are sparse.
  - **"strength"** and **"significance"** are *distinct*!
- **What determines statistical significance?**
  - Standard error! It takes into account the sample size!
  - Measures the accuracy with which the *sample mean* estimates the *population mean*.

Do Two Distributions Have the Same Mean?

- **A very common question**:
  - one sample may be gathered before some *event*, the other sample — after
    - "event" may be treatment, change of a control parameter, etc.
    - we want to know whether or not it made a difference...
- **First thought**: check how many *standard deviations* one sample mean is from the other.
  - may be a useful thing to know: related to the *importance* or *strength* of the difference in means *if that difference is significant*.
  - says *nothing* about whether the difference is statistically significant.
    - The difference in means may be very small compared to the standard deviation, and yet very significant if the sample is large, or it may be quite large but not significant if the data are sparse.
  - **"strength"** and **"significance"** are *distinct*!
- **What determines statistical significance?**
  - Standard error! It takes into account the sample size!
  - Measures the accuracy with which the *sample mean* estimates the *population mean*.
  - Typically,

$$SEM = \frac{\sigma}{\sqrt{N}}$$

Unpaired and Paired Samples

Independent (unpaired) samples

- Two separate *independent, identically distributed* samples from two populations.
- **Example**: two sets of patients — one receives treatment, the other is a control sample.
- May be experimental or observational: e.g., obtain age and gender and check whether the mean ages are different for men and women.

Dependent (paired) samples

- Typically "repeated measures": a group of patients is tested prior to treatment, and then re-tested after the treatment (e.g., is their blood pressure lower than before?).
- Effectively, each patient is his/her own control.
- Random variations between patients are eliminated to a large extent.
- Tradeoff: more measurements are needed, sample size must be doubled.
- Paired samples may be created based on additional variables/parameters. This is used to reduce the influence of confounding factors.

Pooled Variance

- Assume k populations indexed as $i = 1, \ldots, k$
- Assume samples drawn from each population, of size n_i
- *Sample variances* are

$$\sigma_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_j - \overline{x_i})^2$$

- *Pooled variance* is the sum of *sample variances* weighted by d.o.f $= n_i - 1$:

General case:

$$\sigma_p^2 = \frac{\sum_{i=1}^{k} (n_i - 1)\sigma_i^2}{\sum_{i=1}^{k} (n_i - 1)}$$

Equal sample sizes ($n_i = n$):

$$\sigma_p^2 = \frac{1}{k} \sum_{i=1}^{k} \sigma_i^2$$

Equal variances ($\sigma_i^2 = \sigma^2$):

$$\sigma_p^2 = \sigma^2$$

Independent 2-Sample T-Test

- *Test statistics*: the difference in means in terms of weighted SEM:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sigma_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sum_{j=1}^{n_1}(x_j - \overline{X_1})^2 + \sum_{j=1}^{n_2}(x_j - \overline{X_2})^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- Special case $n_1 = n_2 = n$:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}$$

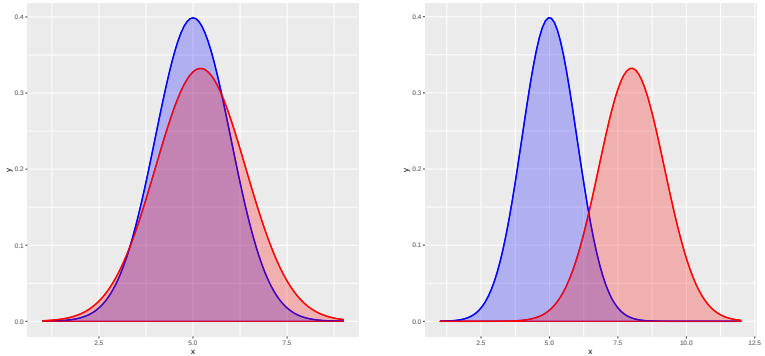
- Special case $\sigma_1^2 = \sigma_2^2 = \sigma^2$:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sigma\sqrt{\frac{2}{n}}}$$

Meaning of the T-Test

- *Test statistics*: the difference in means in terms of weighted SEM:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sigma_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$



- If $|t|$ is "small" then the means are not significantly different, if $|t|$ is "large" they are.
- *Significance* (what "large" and "small" mean):
  - *Null hypothesis* H_0: the means are the same.
  - What is the probability that $|t|$ could be as large or larger than observed *by chance* under H_0?
  - Aha! We got to *hypothesis testing* and *p-value* again!

Student's t Distribution

Probability density function for ν d.o.f.:

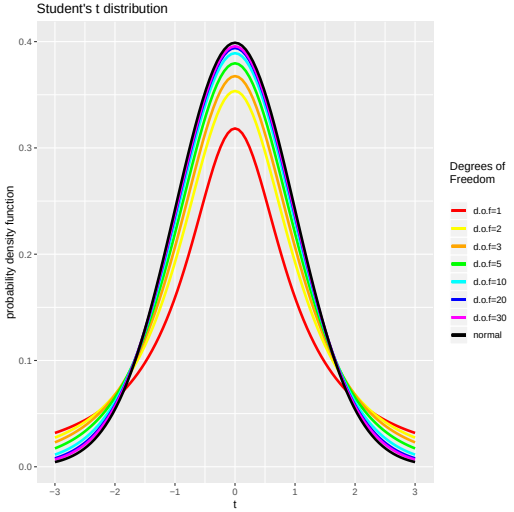
$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

or, equivalently,

$$f(t) = \frac{1}{\sqrt{\nu}B\left(\frac{1}{2},\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

For $t > 0$ and $x(t) = \frac{\nu}{t^2+\nu}$ the CDF is

$$F(t) = \int_{-\infty}^{t} f(t)dt = 1 - \frac{1}{2}I_{x(t)}\left(\frac{\nu}{2},\frac{1}{2}\right)$$



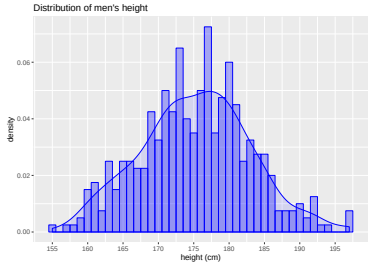
Student's t Distribution in R

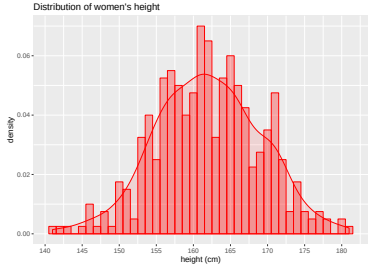
- You don't need to be frightened by the Gamma [$\Gamma()$], Beta [$B()$], and Incomplete Beta [$I()$] *special functions*...
- R has the usual collection of function for the t distribution:

```
dt(x, df) - probability density function (PDF)
pt(q, df) - cumulative distribution function (CDF)
qt(p, df) - quantile function
rt(n, df) - random generator
```

- *Notes:*
  - There is `pt(q,df,lower.tail=FALSE)` for the tail distribution function (the default is `lower.tail=TRUE`).
  - There are other arguments that may be important, but we will ignore them for now.
  - For large number of degrees of freedom (say, $n > 30$, certainly for $n \gg 30$ is very close to the normal distribution.

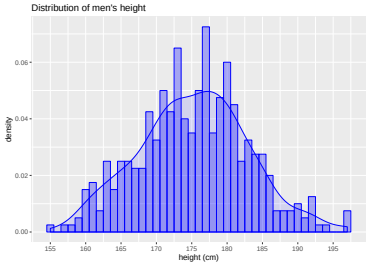
Example: Are Men Taller Than Women?

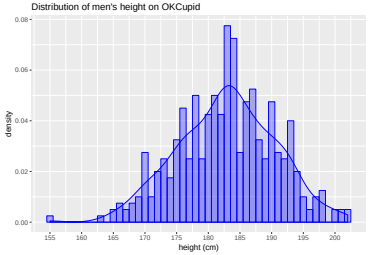





```
> t.test(us.men$height,us.women$height)

Welch Two Sample t-test

data:  us.men$height and us.women$height
t = 25.386, df = 790.74, p-value < 2.2e-16

alternative hypothesis:
 true difference in means is not equal to 0

95 percent confidence interval:
 12.23697 14.28803

sample estimates:
mean of x    mean of y
175.2525     161.9900
```

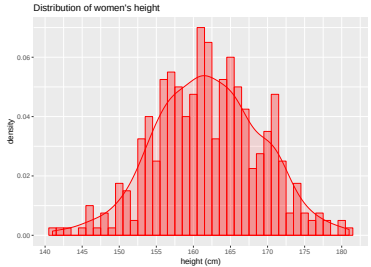
Example: Are Men Taller Than They Are?

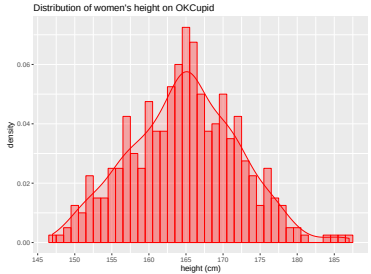





```
> t.test(us.men$height,okcupid.men$height)

Welch Two Sample t-test

data:  us.men$height and okcupid.men$height
t = -13.926, df = 797.99, p-value < 2.2e-16

alternative hypothesis:
 true difference in means is not equal to 0

95 percent confidence interval:
 -8.671231 -6.528769

sample estimates:
mean of x    mean of y
175.2525     182.8525
```

Example: Are Women Taller Than They Are?






```
> t.test(ok.women$height,us.women$height)

Welch Two Sample t-test

data:  ok.women$height and us.women$height
t = 5.5321, df = 797.3, p-value = 4.294e-08

alternative hypothesis:
 true difference in means is not equal to 0

95 percent confidence interval:
 1.800022 3.779978

sample estimates:
mean of x   mean of y
164.78      161.99
```

Student's t-test for Paired Samples

Another HR scenario: hiring committee

We have 2 job candidates evaluated by a hiring committee comprised of 10 members. Each committee member gives each candidate a score from 0 to 10. We want to hire the candidate whose average core is better, *provided that it is significantly better*.

Committee member	1	2	3	4	5	6	7	8	9	10
Candidate 1	6	7	7	6	8	5	9	7	6	6
Candidate 2	7	8	8	7	9	4	9	7	7	7

- The 2nd candidate got higher scores, but the t-test shows that the difference is not significant:

```
> x <- c(6,7,7,6,8,5,9,7,6,6)
> y <- c(7,8,8,7,9,4,9,7,7,7)
> t.test(x,y)
t = -1.0358, df = 17.316, p-value = 0.3146
mean of x mean of y
      6.7       7.3
```

Student's t-test for Paired Samples (Cont.)

- Maybe the significance of the difference in scores is "washed out" by the tendency of some committee members to give higher scores and others to give lower scores? This would *increase variance* and *decrease the significance of the difference in means*!
- Normalize by the *sample standard deviation of the difference in means*:

$$\sigma_{X-Y} = \sqrt{\frac{\sigma_X^2 - 2Cov[X,Y] + \sigma_Y^2}{n}}$$

Applying *paired* t-test:

```
> t.test(x,y,paired=TRUE)
t = -2.7136, df = 9,
    p-value = 0.02386
mean of the differences
                    -0.6
```

Applying *one-sided paired* t-test:

```
> t.test(x,y,paired=TRUE,
         alternative="less")
t = -2.7136, df = 9,
    p-value = 0.01193
mean of the differences
                    -0.6
```

What if t-test Cannot Be Used Directly?

Scenario: switching supplier

Your company receives a component from a supplier. Historically, 2% of the delivered components have been found faulty. The procurement department found a new supplier and an evaluation shipment of 500 units has been received and tested. 18 units have been found faulty. Does the new supplier provide significantly (at 95% confidence level) worse merchandise?

- Approach No. 1:
  - clearly a binomial distribution with mean $P = 18/500 = 0.036$, standard deviation $\sigma = \sqrt{P(1-P)} \approx 0.186$
  - our old friend **"95% confidence interval"** (we use $z_{95\%} \approx 1.96$ for $d.o.f = 499$ — confirm it!)

$$CI_{95\%} = 0.036 \pm 1.96\frac{\sigma}{\sqrt{n}} = 0.036 \pm 1.96\frac{0.186}{\sqrt{500}} \approx 0.036 \pm 0.016$$

  - The historical 0.02 fault rate is marginally inside the CI — inconclusive...

What if t-test Cannot Be Used Directly? (Cont.)

- Approach No. 2 — t-test:
  - But we don't have the samples... or any "historical" data save for the mean fault rate $\mu = 0.02$
  - But we can safely assume that the historical data sample is **much** larger than the $n = 500$ evaluation sample from the new supplier!
  - This means that the contribution of the historical data to the *standard error* is negligible!
  - H_0: the fault rate for the new supplier is 0.02 — just like the historical one
  - The test statistic is

$$t = \frac{0.036 - 0.02}{\sigma/\sqrt{n}} = \frac{0.016}{0.186/\sqrt{500}} = \frac{0.016}{0.00833} \approx 1.92$$

  - *p-value* is the (*one-sided*) tail distribution function for t distribution:

$$p = 1 - \texttt{pt}(1.92, 499) = \texttt{pt}(1.92, 499, \texttt{lower.tail} = \texttt{FALSE}) \approx 0.0277 = 2.77\%$$

  - NB: using *normal* distribution instead of t gives
    $p = 1 - \texttt{pnorm}(1.92) = \texttt{pnorm}(1.92, \texttt{lower.tail} = \texttt{FALSE}) \approx 0.0274 = 2.74\%$
  - At **95% confidence level** the new supplier has a *significantly* higher fault rate...
    They'd better be cheaper...