Lecture 6: The *t*-test

May 17, 2012
GENOME 560, Spring 2012

Su-In Lee, CSE & GS
suinlee@uw.edu

Goals

- The *t*-test
  - Basics on *t*-statistic, confidence interval
  - One-sample *t*-test
  - Two-sample paired and unpaired *t*-test
- R session
  - Doing *t*-test on published gene expression data

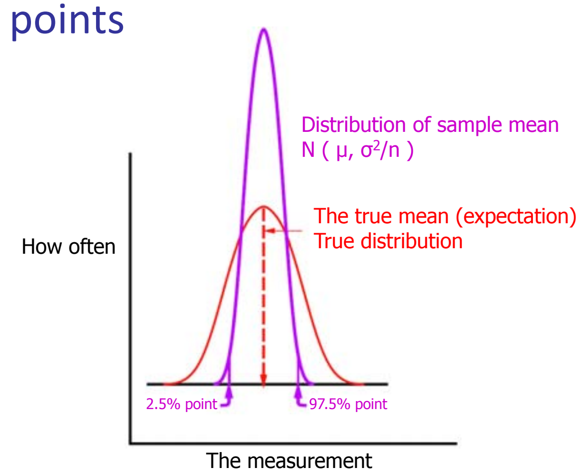
Normality

- Say that $x_1, \ldots, x_n$ are i.i.d. observations from a Gaussian distribution
  - "i.i.d." stands for independent and identically distributed
- Their *sample mean* and *sample variance* are respectively:

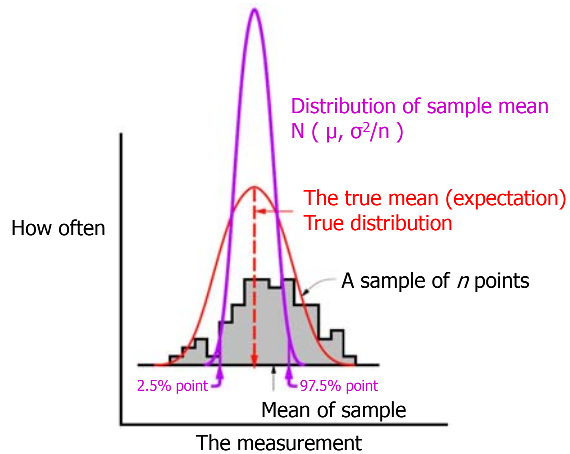
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- How accurate the sample mean will be?

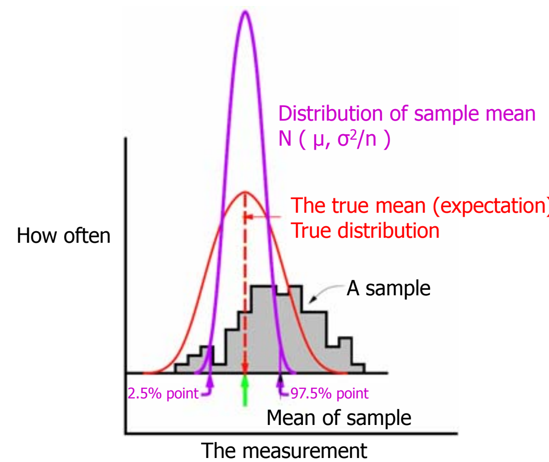
The distribution of the sample mean of *n* points



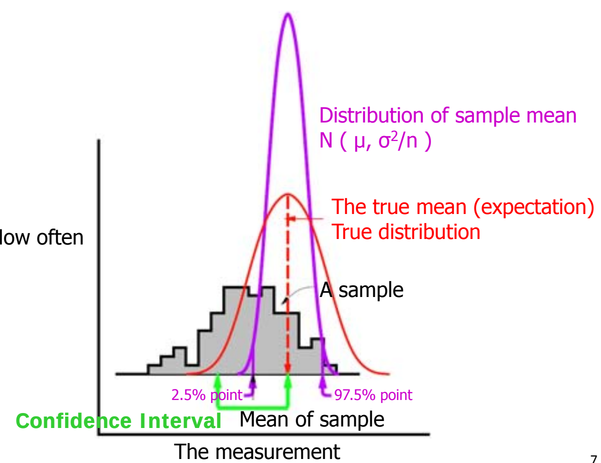
A Particular Sample



Not Any Lower Than This …



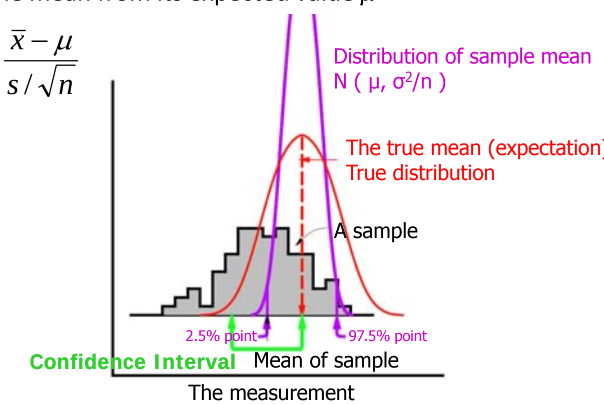
Not Any Higher Than This …



The *t* Statistic

- The number of (estimated) standard deviations of the sample mean from its expected value μ

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$



T-Statistic

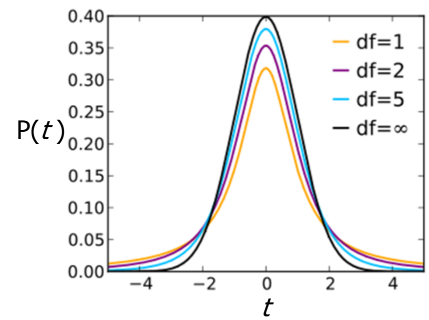
- **Definition:** $t = \dfrac{\overline{x} - \mu}{s / \sqrt{n}}$
  - where, $\overline{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$ $\quad s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- **Interpretation:** The number of (estimated) standard deviations of the *sample mean* from its expected value μ
- The t-value follows a Normal distribution? No
  - We are using sample variance s^2, not true variance
  - Closer to normal distribution as the bigger *n* is.
- The quantity (n-1) is called the *degrees of freedom* of the t value

9

Student's t-Distribution

- The t-values follow the t-Distribution

$$t = \frac{\overline{x} - \mu}{s / \sqrt{n}}$$



  - df = degrees of freedom

10

Student's t-Distribution

W.S. Gosset (1876-1937) was a modest, well-liked Englishman who was a brewer and agricultural statistician for the famous Guinness brewing company in Dublin. It insisted that its employees keep their work secret, so he published under the pseudonym 'Student' the distribution in 1908. This was one of the first results in modern small-sample statistics.

11

One Sample *t*-Test

- Given the following data, assumed to have a normal distribution:

$$x_1,\ x_2,\ \ldots,\ x_n$$

- Hypothesis testing I **Two-sided test**
  - **Null hypothesis H_0:** The mean of the distribution is equal to a specified value μ_0
  - **Alternative hypothesis H_A:** The mean is not equal to μ_0
- Hypothesis testing II **One-sided test**
  - **Null hypothesis H_0:** The mean of the distribution is equal to a specified value μ_0
  - **Alternative hypothesis H_A:** The mean is smaller than μ_0

12

One-Sample Two-Sided Test

- Hypotheses
  - **Null hypothesis H_0:** The mean of the distribution is equal to a specified value μ_0
  - **Alternative hypothesis H_A:** The mean is not equal to μ_0
- Assume that H_0 is true
- Then the *t*-value will have the Student's *t*-distribution

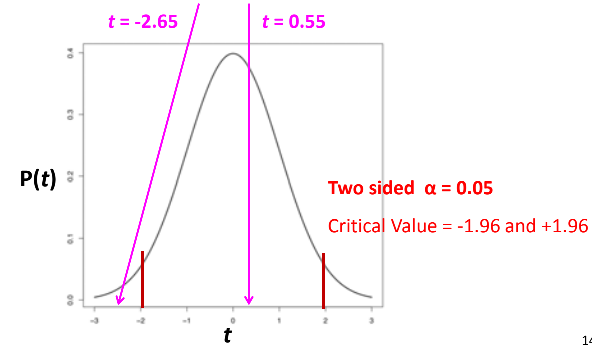
$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

13

One-Sample Two-Sided Test

- You can see how surprising it is to see the *t*-value

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$



14

One-Sample One-Sided Test

- Hypotheses
  - **Null hypothesis H_0:** The mean of the distribution is equal to a specified value μ_0
  - **Alternative hypothesis H_A:** The mean is smaller than μ_0
- Assume that H_0 is true
- Then the *t*-value will have the Student's *t*-distribution

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

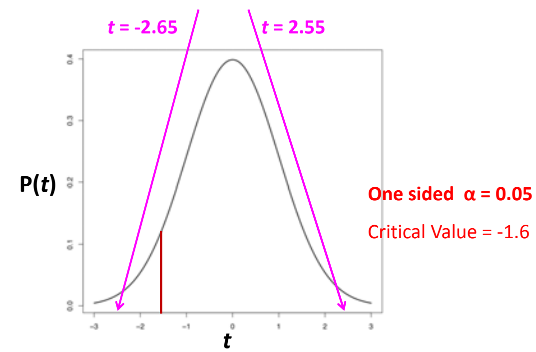
- You can see how surprising it is to see the *t*-value

15

One-Sample One-Sided Test

- **Alternative hypothesis H_A:** The mean is smaller than μ_0

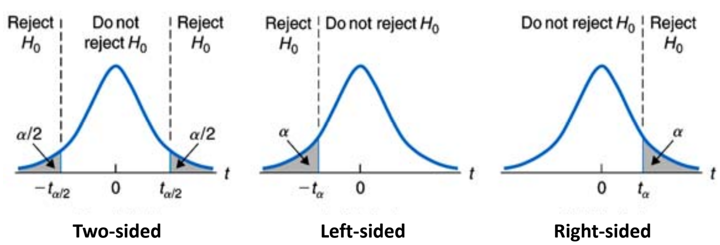
$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$



16

The One-Sample *t*-Test

- **Level of significance, α:** Specified before an experiment to define rejection region



Two Sample t-Test

- **Paired two-sample *t*-test:**
  - There are two samples of the same size (say *n* numbers)
  - The corresponding numbers pair naturally
  - Examples
    - Before-and-after pairs of measurements after giving a drug
    - Expression levels of *n* genes on two samples (one CEU and one YRI)
- **Unpaired two-sample *t*-test:**
  - Two samples might even have different numbers of points (say n_1 and n_2, respectively)
  - There is no natural pair

Paired Two-Sample *t*-Test

- Given the following data (expression levels of *n* genes):

$$x_1,\ x_2,\ \ldots,\ x_n \qquad \text{Before drug treatment}$$

$$y_1,\ y_2,\ \ldots,\ y_n \qquad \text{After drug treatment}$$

- Measure whether the "after" member of the pair is different from the "before" member

$$d_1,\ d_2,\ \ldots,\ d_n$$

After drug treatment
Difference $d_i = x_i - y_i$

- Hypothesis testing
  - **Null hypothesis H_0:** The mean of this sample of differences is 0
  - **Alternative hypothesis H_A:** The mean is not 0
- It is just a one-sample *t*-test of sort we used above

Un-Paired Two-Sample *t*-Test

- Supposed that two samples are drawn independently

$$x_1,\ x_2,\ \ldots,\ x_n$$

$$y_1,\ y_2,\ y_3,\ \ldots,\ y_m$$

  - There is no connection between point 18 from one sample, and point 18 from another
- We want to compare the means of the two samples

Un-Paired Two-Sample *t*-Test

- Supposed that two samples are drawn independently

$x_1, x_2, \ldots, x_n$ ← Assumed to have a normal distribution with mean μ_x

$y_1, y_2, y_3, \ldots, y_m$ ← Assumed to have a normal distribution with mean μ_x

- Is the difference in means that we observe between two groups significant more than we'd expect to see based on chance alone?

- Hypothesis testing
  - **Null hypothesis H_0:** The means of the two samples are equal $\mu_x = \mu_y$
  - **Alternative hypothesis H_A:** Not equal $\mu_x \neq \mu_y$

21

Example hypotheses

- Is there a significant difference between T2D patients and normal people in gene expression level of PPARG (peroxisome proliferator-activated receptor gamma) ?

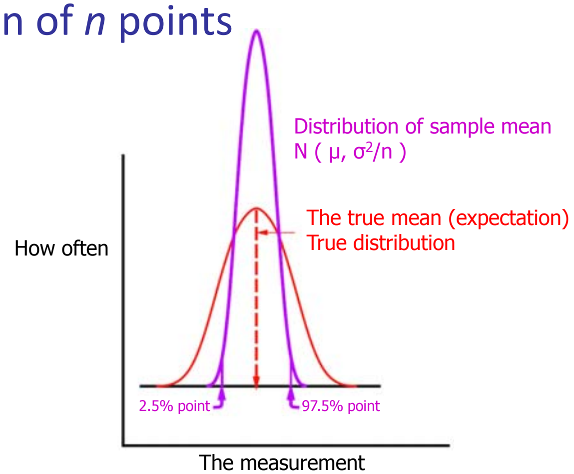
$x_1, x_2, \ldots, x_n$ ← From *n* T2D patients

$y_1, y_2, y_3, \ldots, y_m$ ← From *m* normal people

- Hypothesis testing
  - **Null hypothesis H_0:** No difference $\mu_x = \mu_y$
  - **Alternative hypothesis H_A:** Different! $\mu_x \neq \mu_y$

22

Review: the distribution of the sample mean of *n* points



23

Theoretically…

- The distribution of the sample mean difference, $\bar{x} - \bar{y}$

$$\bar{x} - \bar{y} \sim N(\mu_x - \mu_y, \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}})$$

- Let's think about how the t-value should be defined here

$$t = \frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$$

Under the null hypothesis, $\mu_x = \mu_y$

- As before, we usually have to use the *sample variance* because we don't know the true variance

- So, again becomes a *t*-distribution, not a normal distribution

24

Un-Pooled Variances

- Just replace the true variances with sample variances

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

- The t-statistic has the Student's t-distribution with degrees of freedom *v*
- It is complicated to figure out *v* here!
- A good approximation is given as ≈ harmonic mean $\frac{2}{\frac{1}{n} + \frac{1}{m}}$

Pooled Variances

- If you assume that the variance is the same in both groups, you can pool all the data to estimate a common variance.
- This maximizes your degrees of freedom (and thus your power)
- The *t*-statistic is then defined as:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}} = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Pooled Variance

- Pooling variances:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \Leftrightarrow (n-1)s_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s_y^2 = \frac{\sum_{i=1}^{m}(y_i - \bar{y})^2}{m-1} \Leftrightarrow (m-1)s_y^2 = \sum_{i=1}^{m}(y_i - \bar{y})^2$$

$$\therefore s_p^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{m}(y_i - \bar{y})^2}{n+m-2}$$

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

Degrees of freedom

Un-Paired Two-Sample *t*-Test

- Hypothesis testing
  - **Null hypothesis H_0:** No difference $\mu_x = \mu_y$
  - **Alternative hypothesis H_A:** Different! $\mu_x \neq \mu_y$
- ***t*-statistic:** $t = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$
  - where $s_p^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{m}(y_i - \bar{y})^2}{(n-1)+(m-1)}$
  - The *degrees of freedom* is (n+m-2)