Handout 1

About Your TA

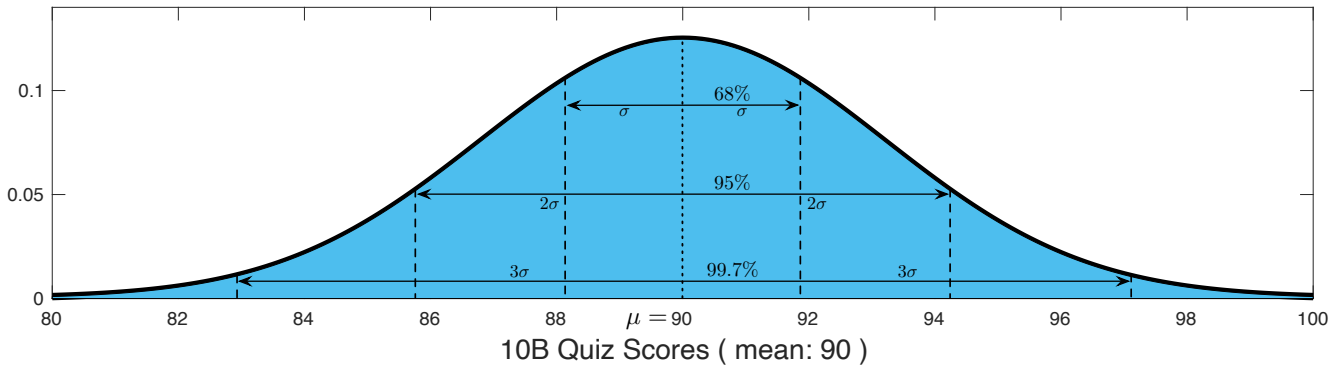
- **Haoche (Howard) Hsu**
- Lab Sections:
  1. Lab-6 Wed. 12-1 pm (MSTB 226)
  2. Lab-7 Thur. 9-10 am (ALP 3610)
- Office Hour: Tue. 3:30-5:30 pm (SSPA 3182)
- My Website: http://www.haochehsu.com (Handout can be found at the *Teaching* section)
- Email: haoche.hsu@uci.edu

1 Logistics

- Homework 8 is due week 4 (1/29, 1/30 in lab)
- Quiz 1: Thur. 1/23 (in class)
- Lowest 3 lab scores will be dropped
- Strict homework submission format requirements
- Lab sign-in & sign-out
- Any comments feel free to use the anonymous *Feedback Survey* (on the website)

2 The Bell-shaped curve

Before we review *Hypothesis Testing*, we need to spend some time to discuss the following bell-shape curve.



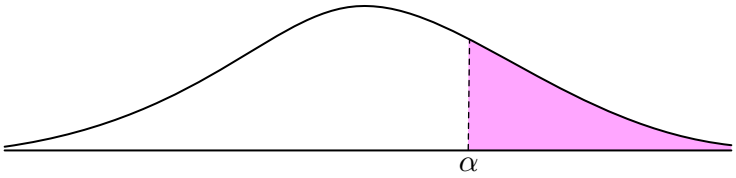
The bell-shaped curve characterized a distribution that values near the center (mean) will appear more frequently than values at the tails. For instance, if the 10B quiz scores follow a bell-shaped distribution centering at a score of 90, then it means that most of the people get scores like 88, 89, 91. Even though someone might score a 100 but the number of students who get full credit is very small since observing the graph, 100 is located at the tail of the curve. People have study this curve for years and derive a lot of useful properties. This bell-shaped curve is so famous that we call it the *Normal Distribution.*

Many Mathematicians such as Gauss, Adrain, and Laplace have contributed to the study of normal distribution. This distribution is also known as *Gaussian Distribution*. If we provide a mean (μ) and a variance (σ^2), then the normal distribution can be illustrated by the following function $f(x|\mu, \sigma^2)$:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}}.$$

So if we plot this function, we will obtain the bell-shape curve. The normal distribution has some useful properties:

- It is **centered at the mean** where the peak is.
- The distribution is **symmetric**. If you slice the distribution from the center then the curve on each side has a perfectly symmetrical shape. The left area under the curve will be the same as the right.
- The middle point of the distribution is the point with the highest frequency. This **midpoint** is the **mean**, the **median**, and the **mode**.



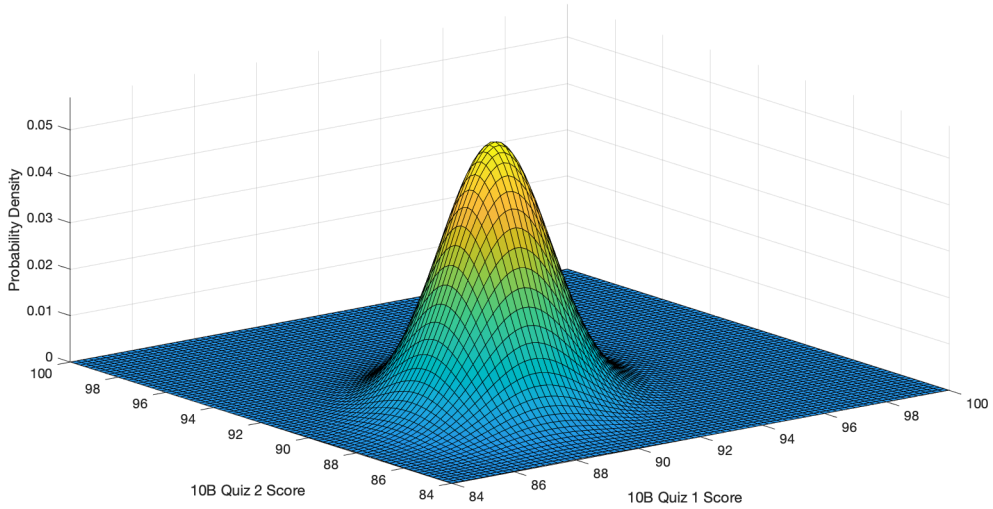
With these properties, it is easy to calculate the shaded area (graph above) if you give me a α. Furthermore, the normal distribution contains a lot of information. For example, if you get 86 on the quiz and the mean (μ) is 90 and the standard deviation (σ) is 3, then if the test result follows a normal distribution, referring to the graph in page 1, you can conclude that around $50\% + (68\%/2) = 84\%$ of people's score in the class is higher than you.

Therefore, when we do statistical inference or calculate probabilities, we like to associate things with the normal distribution. For example, when we write down a structural model to analyze consumer demand, we often assume the data (people's preference) is normally distributed. Most importantly, we will try to use normal distribution when conducting hypothesis testing.

Just a side note, normal distribution does not only exists in 2-dimension. Suppose the first and second quiz scores both follow a normal distribution and these two scores affect each other in a sense that if you understand the materials well enough for the first quiz, it is highly likely that you will get a high score in the second quiz, then the quiz scores can be described by the following function:

$$f(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$$

where Σ is the covariance matrix. This is the *multivariate* normal distribution. I have plotted this normal distribution for you to visualize in a 3-dimension graph:



3 Hypothesis testing in a nutshell

The power of statistics kicks in when we want to know some important aspects of the population (the data generating process, DGP) but we only observe some sample data. Hypothesis testing consists of several steps. I will construct a fictitious example to demonstrate the inference process.

3.1 The problem

Suppose this year as Apple debuts their iPhone 11, they decide to **collect the consumer's age** when a purchase is made. So in Apple's database, every iPhone 11's serial number is associated with a buyer's age. And months after the lunch of the new iPhone, Apple claims that **the average age of iPhone 11 buyers is 18**. But I am not convinced, I think the company may have made some calculation errors. It is also possible that they might forget to collect the age for some sales.

3.2 Obtain data

To investigate this problem, I go to the *South Coast Plaza* and randomly survey 5 people who have an iPhone 11 in hand and obtain the following data:

iPhone Model	Color	Age	Gender	Capacity	Carrier
11	White	30	female	256 GB	Verizon
11	(PRODUCT)RED	49	male	128 GB	Verizon
11	Green	55	male	64 GB	AT&T
11	Green	42	female	64 GB	Verizon
11	Purple	70	male	128 GB	T-Mobile

3.3 State the hypotheses

To begin my inference, I will need to state the hypotheses[1] **based on my problem**:

$$H_0 \text{ (null hypothesis)} : \mu = 18$$
$$H_1 \text{ (alternative hypothesis)} : \mu \neq 18$$

Notice that no matter what kind of hypothesis you are making, the **null hypothesis always contains the equal sign** because H_0 is the statement we are challenging. It is the "standard"of the test, the statement we assume to be true if there is no further information. Then we try to find evidence that supports the alternative hypothesis ($H1$). The null hypothesis (H_0) is rejected if and only if the supporting evidence for H_1 is *significant*.

How do we find the *supporting evidence* for H_1? We look at the probability of H_0.

3.4 Testing

Recall that H_0 is our standard. We **treat H_0 as the truth** as we search for evidence:

If H_0 is true, then no matter how many samples collected,
those age numbers are drawn from a distribution with mean 18.

This means that if what Apple has claimed is true, the average age of all iPhone 11 buyers is 18, then regardless where I collect the data, at *South Coast Plaza* or at *Spectrum Center*, it will be more likely for the surveyed age data to be $\{15, 20, 25, 19, ...\}$ since numbers around the mean should appear more frequently. If the age of the five people I've asked often consists of numbers such as $\{40, 50, 71, 38, 66\}$, then it is fair to guess that the true average age should be some number much higher than 18 years old.

[1] H_0 can be either: $=$, $\leq$, $\geq$. H_1 can be either: $\neq$ (two-tailed test), $<$ (left-tailed test), $>$ (right-tailed test).

Looking at the **sample** data that I've collected in section 3.2, the age of the five people are $\{30, 49, 55, 42, 70\}$ with **sample mean 49.2** years old and **sample standard deviation 14.889**. Can we directly compare 49.2 to 18 and make any decisions? No! Notice that the sample mean is affected by the variance of the sampling distribution. Hence, we need to **normalize the sample mean** before making a comparison.

3.4.1 Calculate the test statistic

We will use test statistics: *z-score* or *t-score* to normalize our sample mean. Since there are only 5 observation[2], we will use t-score. The t-score normalization is the following data transformation:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\text{SE}} \quad \text{where} \quad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\boldsymbol{n-1}}} \qquad (s \text{ is also denoted as } \widehat{\sigma})$$

We want to measure how far is the sample mean to 18 (the H_0) by subtracting the sample mean with the true mean (since we currently assume H_0 is true) and divided by $\frac{s}{\sqrt{n}}$ where s is the sample standard deviation and n is the number of observations. This t-score (t) is also known as "t statistic"or "obtain t."

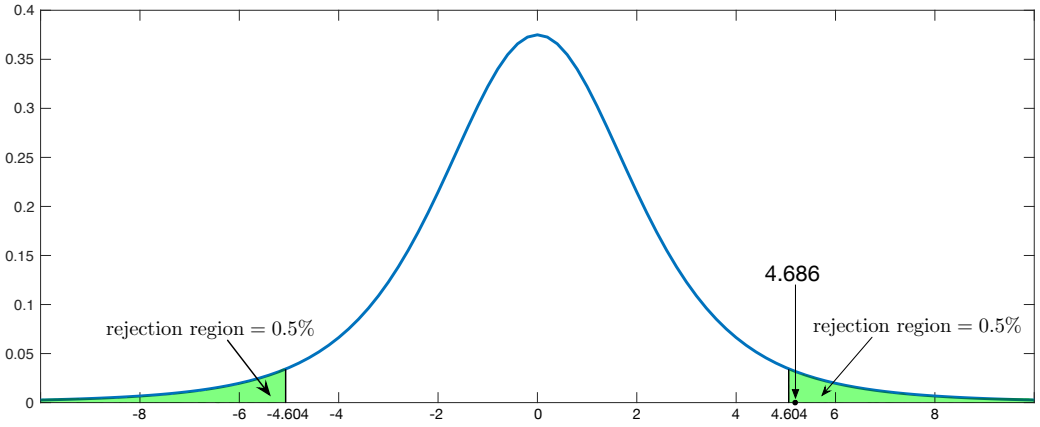
Moreover, the reason we choose this equation to normalize our sample mean is that this formula, by the *Central Limit Theorem*, will converge to a (standard) normal distribution which is something that we can take advantage of. Plug in the data, we "obtain"our t statistic:

$$\frac{49.2 - 18}{\frac{14.889}{\sqrt{5}}} = 4.686$$

Now we obtain the normalized sample mean 4.686, we are ready to make a decision.

3.4.2 Construct rejection regions

The t-score follows a t distribution which has a bell-shaped curve. To be more precise, since the statistic is calculated with 5 observations, this *t-score assuming H_0 is true* will follow a *t distribution with 4 degrees of freedom*:



Let's observe this t distribution. We can see that a 0 t-score has a high probability to occur, at-score of -6 is less likely to occur and at-score of 8 is, even more, less likely to occur. Also, notice that the area under the curve is probability. Since probabilities sums to 1, the entire area under the bell-shaped curve is 1. Next, let's look at some numbers. Take 4.604 as an example, the area under the curve above 4.604 (the shaded area on the right) is 0.005. That means the probability for us to *draw a number from this t distribution that is larger then 4.604* is 0.005, a really low chance. Similarly, the probability to draw a number from this distribution which is smaller than -4.604 is also 0.005. Finally, the probability to draw a number which is greater than 0 is $\frac{1}{2}$.

Before we move on to testing, we need to specify some "features"about this test:

We will conduct a test with 99% confidence level.

[2] Choice of test statistics: when observations is ≥ 30, we use z-score. When observations is ≤ 29, we use t-score.

A 99% confidence level test means that for all the decisions made from this test, 99% is correct. Only 1% chance that we will make a mistake (i.e. cause a *type I* or *type II* error). These mistakes are caused by **making wrong rejections**.

Furthermore, since the confidence level of our test is 99%, the significance level (α) is 1%. This implies that the rejection region has area 0.01. Since our alternative hypothesis (H_1) is "not equal $\neq$", this test is also a two-tailed test. So the rejection region (area under the curve) on each side of the tail has area 0.005.

Next, we need to find the critical value. The critical value is the t-score that will make area of the tail region (area under the curve above the critical value) 0.005. Since this is a two-tailed test, we have two critical values: 4.686 and -4.686 obtained from the Student's t distribution table[3].

Why do we make rejections? Recall that we do all these steps while making the assumption that H_0 is true[4]. So given H_0, we look at the t distribution with $n = 5$ ($df = 4$) on previous page. The probability for a t-score that is larger than 4.604 **or** smaller than -4.604 to appear is 0.01. If H_0 is true, then the t-score of our 5-observation sample should be between -4.604 and 4.604 which has a larger probability to occur. For example, if our sample mean is exactly 18, then after the normalization, we will obtain a t-score of 0 which is the peak of the t distribution. This means that given H_0, a data with sample mean 18 is highly likely to occur.

In section 3.4.1, we have already calculated the t-score of our sample is 4.686. But from the t distribution assuming H_0 is true with 4 degrees of freedom, the probability for the sample to have a t-score like this is 0.005. This indicates that the 5 sample data obtained at the South Coast Plaza is **very unlikely** to be drawn from a distribution with mean 18.

3.5 Draw a conclusion

The fact that our sample doesn't support the H_0 assumption is the significant evidence supporting H_1. Hence, we choose to reject H_0 even though there might be 0.01 of chance that this is a mistake.

Even though there are other approaches such as *Bootstrapping* to estimate the true population mean, hypothesis testing is still an elegant and simple method to make inference from observed sample.

[3] The critical values for a two-tailed test with t test statistic, 99% confidence level, 4 degrees of freedom is 4.604. Since t distribution is symmetric, -4.604 will also generate a tail region with area 0.005 on the left side.

[4] That's why when we normalize the sample mean, we deduct it by 18, the assumed true mean.