Week 1 Tuesday Hr 2 (Review 1)

- Samples and Populations

- Descriptive and Inferential Statistics

- Normal and T distributions

One of the primary goals of statistics is to make **statistical inferences** on of a **population**. A population is a blanket term for anything that is too large or difficult to measure directly.

Population Examples:

- All the people living in Burnaby.
- All the water in Deer Lake.
- All the inpatients at VGH.

However, in a statistical sense, populations must be a well defined set of the same type of thing.

Population Non-Examples:

- The city of Burnaby (what in the city? People? Cars?),
- All the water around (around where? One lake? In the sky?).

See Reading Note 1.1: generalizations.

Statistical inferences are made about **parameters**, which is some numeric variable of interest relating to the population.

Parameter examples:

- The average concentration of pollutant in Deer Lake (measured as ppm or mg/L),
- The average days that VGH inpatients have been admitted to the hospital,
- The proportion of VGH inpatients that have signs of pneumonia.

Parameters also need to be well-defined variables.

Parameter non-example:

- How sick are people at VGH right now (‘sick’ is ill-defined).
- Whether or not the patient in VGH Room 1234 has cancer. (does not pertain to a population).

Parameters are very difficult or impossible to measure directly because they involve the whole population. However, we can take a **sample** of a population and directly measure a **statistic** from that.

To repeat:

***S*tatistics describe *S*amples**

***P*arameters describe *P*opulations**

There are many ways to take a sample (See Reading note 1.2, sampling), but in this class we will assume that every sample is a **simple random** sample (SRS) unless it is stated otherwise.

In a simple random sample, each member of the population has an **equal chance** of being selected, and every possible sample has an equal chance of being selected.

Example 1, Part 1: If the parameter of interest is ‘proportion with pneumonia signs’ and the population is ‘inpatients at VGH’, we can take a **simple random sample (SRS)** of 10 of the inpatients.

If we find that 6/10 of the patents have signs of pneumonia, then we could make an inference that 0.60 of all the inpatients have signs of pneumonia.

In Example 1, we inferred using the sample proportion p,

$$p = \frac{x}{n}$$

…where ***n*** is the **sample size**. In this example ***n*** is the number of patients that were tested.

x is the number of sample units that match the feature of interest. In this example ***x*** is the number of tested patients who showed signs of pneumonia.

Example 2, Part 1: We take a simple random sample (SRS) of 5 locations in Deer Lake, and collect 1 Litre of water from each.

The total pollutant in all five litres is 12 mg. The mean concentration of pollutant across the samples is 2.4 mg/L.

We make a statistical inference that the concentration of pollutant in all the water of the lake is 2.4 mg/L.

In Example 2, we inferred using the sample mean $\overline{x}$, called x-bar,

$$\overline{x} = \frac{\sum x}{n}$$

where ***x*** is the individual measurements and **Σ** is the summation sign, meaning ‘add these up’.

Therefore, **Σx** means the total of the measurements. In this example, that’s the total of 12mg that we found in 5 Litres.

$$\bar{x} = \frac{\sum x}{n}$$

n, as always, is the **sample size**. In this case, we took samples from 5 different locations, so ***n=5***. Just like in Chemistry, if there are units, we preserve them. The amount sampled was in Litres, and the amount of pollutant was in mg, so the sample mean $\bar{x}$ is in mg/L.

The value that we are inferring about in each example is a **parameter**.

With most (possibly all) of the statistics we will be computing in this course, the sample statistic has a population parameter that it estimates directly. More technically, each statistic we will look at is an **unbiased estimator** of some parameter.

In Example 1, the sample proportion (p = 0.60) is an unbiased estimator of population proportion (π = 0.60).

In Example 2, the sample mean ($\bar{X}$ = 2.4) estimates the population mean (μ = 2.4) without any bias. We don’t and can’t know what the true values for the parameters are, but these are our best guess.

Note that statistics are traditionally labelled with regular (Latin) letters and parameters are labelled with Greek letters.

You could directly find a parameter value, but it would take…

something drastic.

Descriptive and Inferential Statistics

Another primary goal of statistics is to **describe samples**. Often we do both **descriptive** and **inferential** statistics in the same analysis.

Measures of **central tendency (**also called **location)** describe where the middle or centre of the data is, for different definitions of ‘centre’.

Measures of **spread** (also called **scale**) describe how much the numbers in a sample differ from each other.

With an appropriate measure of central tendency and measure of spread, we know enough about a sample to both estimate the value of a parameter AND have an idea of how precise that estimate is.

The most common spread/scale measure is the **standard error**.

If we are estimating a proportion, it’s called the standard error of **the proportion**.

If we are estimating a mean, it’s called the standard error of **the mean**.

Example 2, Part 2: Between the 5 sample locations where we took water from Deer Lake, the mean pollutant concentration is 2.4 mg/L **and the standard deviation is 1.1 mg/L**.

We also know from previous testing of the lake that measurements are **normally distributed**.

Therefore our estimate of the mean pollutant across the whole lake is 2.4 mg/L with a **standard error** of 0.49 mg/L

(standard error = s / sqrt(n) = 1.1 mg/L / sqrt(5)) .

To repeat:

Standard Deviations describe the spread of **individual observations.**

Standard Errors describe the spread of **parameter estimates.**

You can’t change a standard deviation. It’s a property of the population. However, you CAN make a standard error smaller taking additional observations (getting a bigger ***n***).

The sample proportion and sample mean are just two of many statistics that can be used to describe a sample.

Other measures include the median and other quantiles, the skew, and the standard deviation.

Statistic	**Description**	**Formula**
Mean, x-bar	Measure of central tendency, sensitive to extreme values (outliers).	$\bar{x} = \frac{\sum x}{n}$
Median, m, Q2	Measure of central tendency, robust to extreme values (outliers). Also a quantile measure.	No formula. Find the value(s) with 50% of the data points below/above.
Lower Quartile, Q1	Quantile measure	Find the value(s) with 25% of the data points below it
Lower Quartile, Q3	Quantile measure	Find the value(s) with 75% of the data points below it

Statistic	**Description**	**Formula**
1st , 10th, 20th, 30th … 99th percentile, q_1, q_{10}, q_{20}, q_{30}, … q_{99}	Quantile measures	Find the value with 1%, 10%, 20%, 30%, … 99% of the data points below it.
Proportion, p	Self explanatory. Technically a measure of central tendency.	$p = \frac{x}{n}$

Statistic	**Description**	**Formula**
Standard Deviation	Measure of spread. Sensitive to extreme values.	$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2}.$
Variance	Measure of spread. Standard deviation squared.	$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2.$
Skewness	Shape statistic. Measures the direction and strength of the skew of a unimodal distribution	Beyond the scope of this course.

Statistic	**Description**	**Formula**
Chi-Squared Statistic	Goodness-of-fit statistic. Describes how well a set of category responses fit some pre-defined distribution.	$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$
Shapiro-Wilks Statistic	Goodness-of-fit statistic. Used to test if a set of numbers fits the normal distribution.	Beyond the scope of this course

We won’t be computing the Shapiro-Wilks statistic by hand in this class, but the name is important to know.

Some of the computer output we will examine in the diagnostics section of this course will include the Shapiro-Wilks test.

This is a test to determine if a data set could be considered normally distributed. Normality is a major assumption for many methods, so testing for it is useful.

Shapiro-Wilks,

not to be confused with Sharpie Walks.

Normal and T distributions

In Example 2, Part 2 we assumed that the distribution of the lake measurements was normal (also called the Gaussian distribution, or z distribution).

This is an important assumption because… (IMPORTANT)

If the distribution of a single measurement is

Normal(μ, σ) ,

Then the distribution of the mean of n measurements is

Normal(μ, $\sigma/\sqrt{n}$) .

Where Normal(A, B) denotes a normal distribution with a mean of A and standard deviation/error of B.

In the case of Example 2, Part 2, the property would read:

Since the distribution of a single measurement is

Normal (2.4, 1.1) ,

The distribution of the mean of 5 measurements is

Normal (2.4, 1.1 / $\sqrt{5}$) =

Normal(2.4, 0.4919)

The normal distribution is extremely important. When data is normal, all sorts of other assumptions can be made about it.

It’s also an abstract ideal – no real data is perfectly normally distributed, but lots of data is close enough that we can assume that it is.

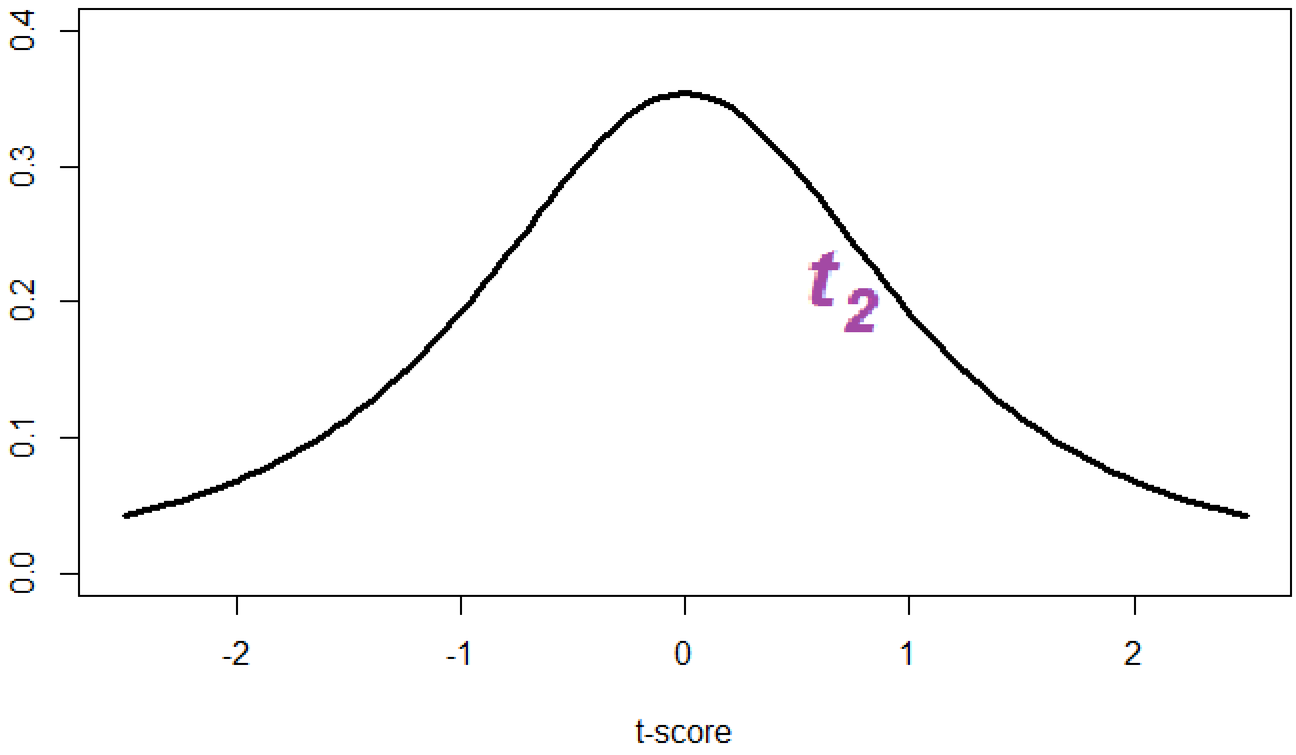
In the normal distribution, we assume that the true (i.e. population) standard deviation sigma is known exactly.

However, since sigma is a parameter, we would need to measure the entire population to know it.

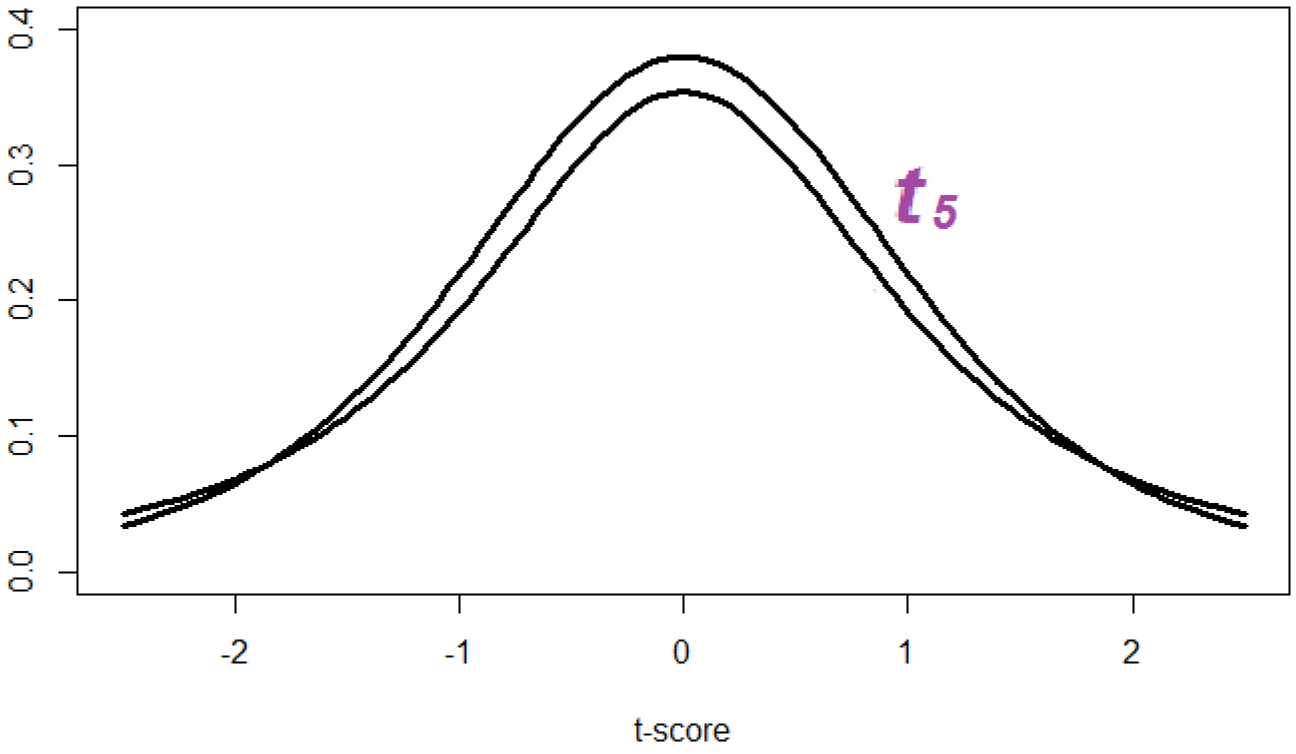
More realistically, Student’s T distribution only assumes that you can estimate the standard deviation with its statistic, ***s***.

The T distribution behaves almost exactly like the normal.

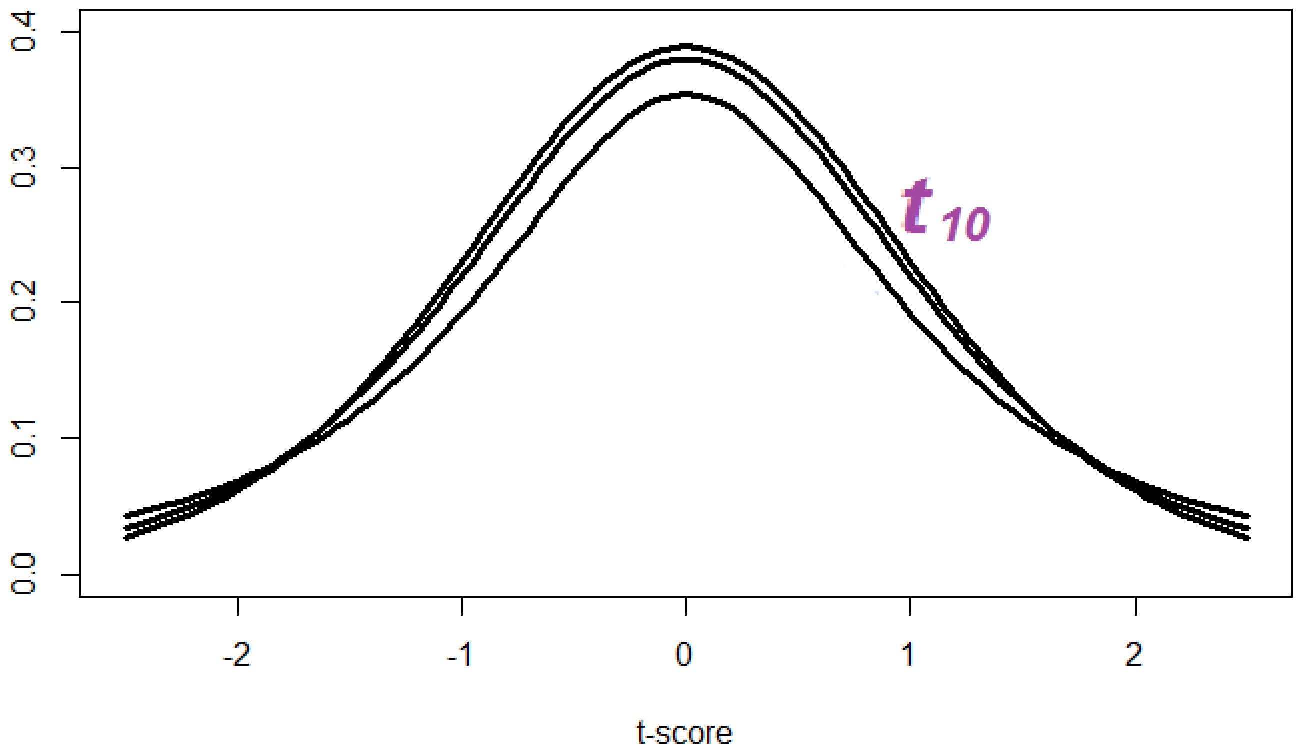
Student’s T distribution is wider and more diffuse/spread than the normal distribution to account for the **additional uncertainty surrounding** σ**, the standard deviation**.



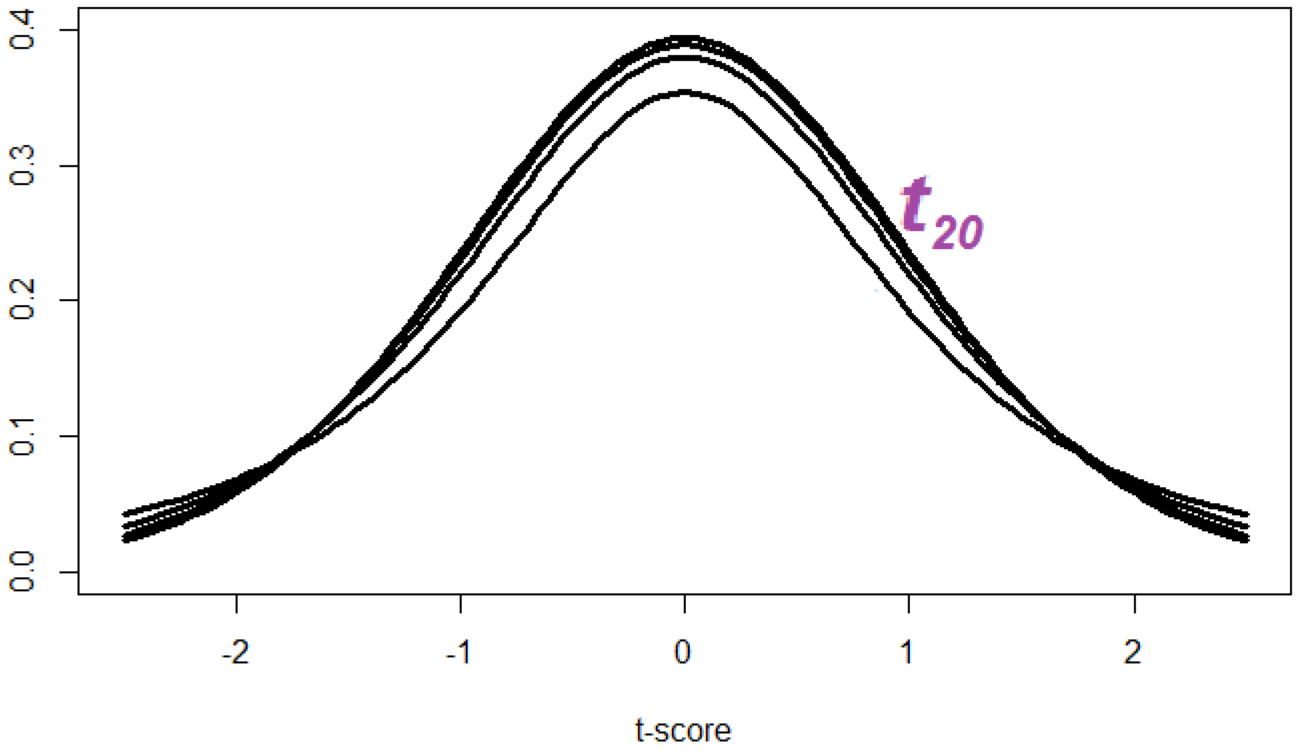
The T distribution also uses **degrees of freedom**, or df. For the T distribution, these degrees of freedom represent the amount of information about the standard deviation that we have from our sample, or samples.



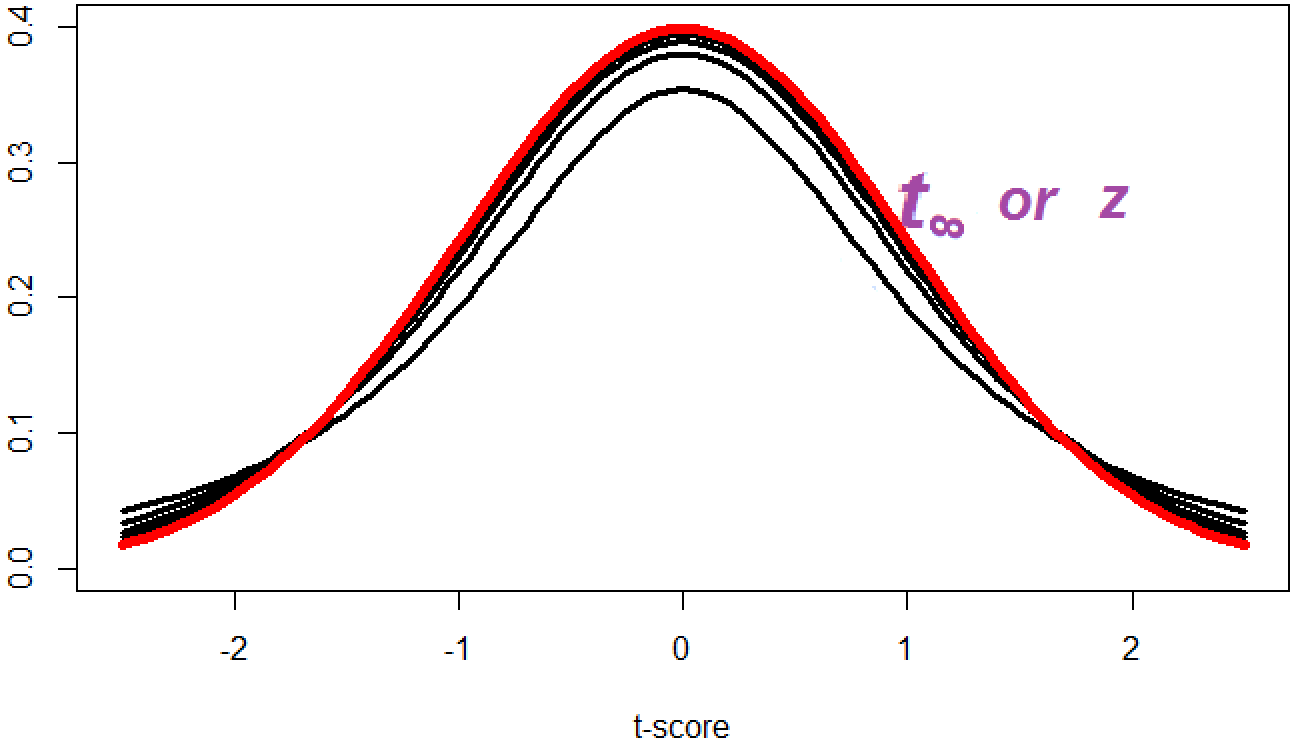
For the T-distribution, the degrees of freedom is determined by the **sample size** (or **sample sizes**, in the two-sample case). There are some complications, but all-else-being-equal, more observations means more degrees of freedom.



As the number of degrees of freedom **increase**, the information about the standard deviation improves, and the estimate ***s*** becomes more accurate and precise.



When there are many degrees of freedom (i.e. when df is large), the estimate of ***s*** is so good that we can reasonably act as if we know the exact value of sigma (i.e. use the Normal distribution).



That’s why the normal distribution (i.e. the z distribution) shows up on the t-table as “t with infinity df”.

There is a mathematical reason behind the T-distribution’s trend towards normality when df is large; the Central Limit Theorem.

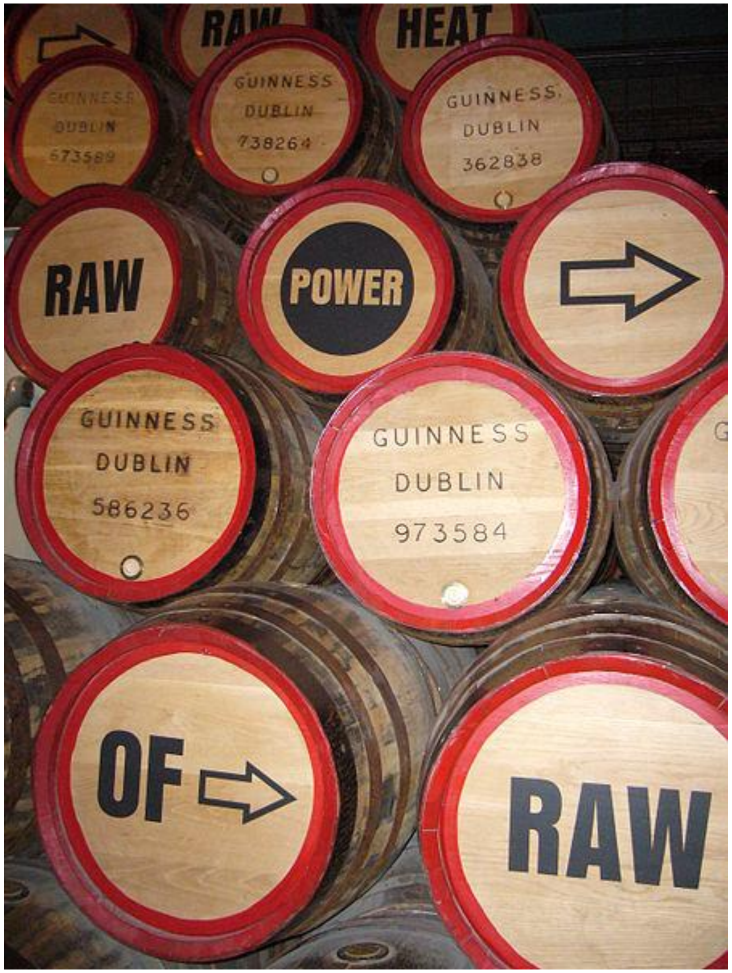
IMPORTANT!!!

Central Limit Theorem: ***The sum or mean of many independent observations from the same population tends towards a normal distribution.***

The T-distribution is used to describe either a single mean or a difference between two means.

We assume that each of the sample numbers that goes into that mean (or means) is independent.

A large part of what follows in this class will also be estimating means of different kinds, so the Central Limit Theorem and normality will come back several times in this course.



True story: The first use of the T-distribution was in selecting varieties of barley for Guinness beer. (see: William Gosset)

Break image sources:

1. Hades from Disney’s Hercules.

2. Shoes from hearttreehome.com

3. Wikimedia user Sebb.

On Thursday:

- Binomial and F-Distribution distributions

- Hypothesis Testing

- Confidence Intervals

***Reading Note 1.1 (For interest only, not on a test)**:* Some populations are dynamic. That means they change over time. People move into and out of Burnaby, and Deer Lake fills and empties with the seasons. How might you handle or work around this problem?

Short answer: Pick a specific time (e.g. Deer Lake at 4:00 pm, April 11, 2015).

Long answer: This is a major problem that gets solved in a lot of different ways in different situations. You might take measurements at regular intervals and make inferences about the trend of some variable of interest. This is called a time-series analysis, and it’s used for looking at the trends of environments over seasons, cities over years, and corporate stock prices over the course of a day. Epidemiologists use data that changes by time and space to infer if a disease is spreading.

There are other fields besides time-series that look at dynamic data or changing data over time and space. For example, there are the fields of spatial data and survival data. We won’t be covering these fields in this course; the point this note is to highlight that there are a lot of generalizations and complications that can be applied to problems in statistics. A major focus of this course, for example, is generalizations of linear regression into a wider range of situations.

Reading Note 1.2 (For interest only, not on a test): Simple random sampling (SRS) is just one of many sampling options. Other options are involve mathematical complications, but have the their own practical advantages. These methods include…

- Non-random/convenience samples: sampling the most convenient members of the population. This is the easiest, but least statistically valid method.

- Stratified samples / two-level sampling: Splitting up the population into groups/strata first, sampling random groups, and using SRS without each group. Sometimes easier than SRS and with similar results, especially when sampling over a large geographic area.

- Quota sampling: Choosing beforehand the number of people from each group you want, and using SRS until that number from each group is met. Usually, for cost reasons, observations beyond each group’s quota are rejected.

- Systematic samples: Taking observation from a population over time at fixed intervals (e.g. at 11am of each day), used to account for non-independence between observations that are near in time.

- Snowball / Network / Recruitment / Respondent driven samples: Selecting a small ‘seed’ group of a population and having members of that seed group recruit new people into the sample. Useful when a group is difficult to find, but well connected, such the homeless or endangered species.

- Transect samples: Travelling along a selected geographical path (i.e. a transect) and taking every observed population member along the path as the sample. Useful in field-based sciences such as forestry and environmental science.