Normal distributions and the Central Limit Theorem

EDUC 641: Unit 3 Part 3

David D. Liebowitz

UNIVERSITY OF OREGON

Roadmap

Research is a partnership of questions and data		What types of data are collected?	
		Categorical data	**Continuous data**
What kinds of questions can be asked of those data?	**Descriptive questions**	• How many members of class have black hair? • What proportion of the class attends full-time?	• How tall are class members, on average • How many hours per week do class members report studying, on average?
	Relational questions	• Are male-identifying students more likely to study part-time? • Are PrevSci PhD students more likely to be female-identifying?	• Do people who say they study for more hours also think they'll finish their doctorate earlier? • Are computer-literate students less anxious about statistics?

Class goals

- Describe special features of a normal (and standard normal) distribution
- Interpret a z-statistic table
- Describe the distribution of repeated sample statistics drawn from a population, how this relates to the Central Limit Theorem (CLT) and how this is informative to statistical hypothesis testing
- Determine whether the mean value of a sample is different than a defined population mean, both when the population standard deviation of the variable is known (z-test) and when it is unknown (one-sample t-test)

Life expectancy data

Reminder of our motivation:

Suppose you are working for the World Health Organization and are investigating life expectancy across different regions. Using this dataset, we can ask questions like:

> - How does life expectancy compare in high- vs. middle- and low-income countries?
> - Do individuals living in countries with more total years of attendance in school experience, on average, higher life expectancy?

But, before making comparisons between social/economic conditions, we want to start with describing our data. Let's say we *know* information on life expectancy about all the countries in the world, but we can only collect information on things like school experience on *some* countries.[1]

Our second task: How similar are these subset of countries' average life expectancy to our known "population" average?

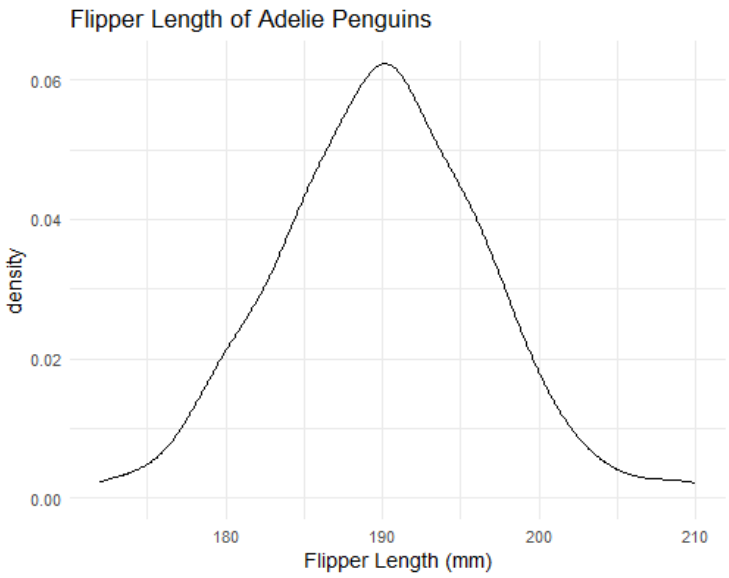
Note: These next few classes are focused on concepts rather than programming, so focus your energies there.

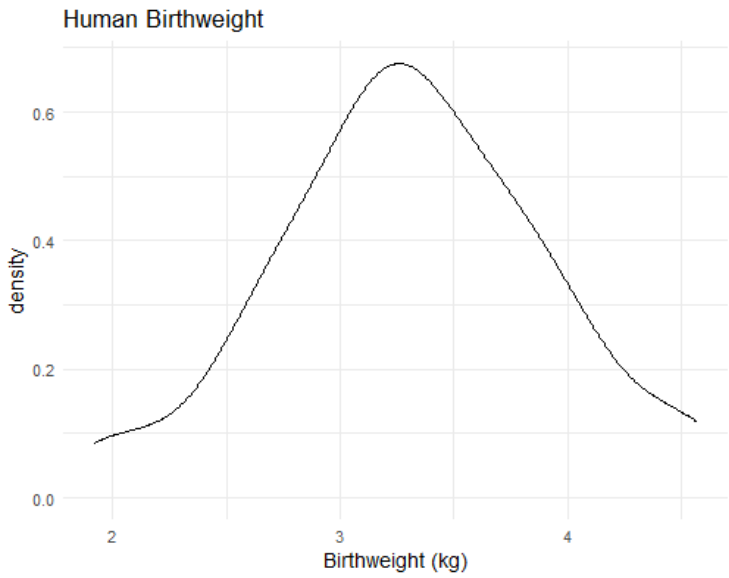
[1] This is somewhat of a simplification. As we will learn, even if we had data on *all* countries, we wouldn't really know the "population" average; in fact, we almost never do and we are always estimating it.

Normal distributions

Bell curves

You have likely heard of a frequent phenomenon that occurs when we quantify observations about the social and natural worlds: most observations tend to cluster in the middle, with others drifting out into narrower tails. You may have heard these described as "bell curves."





Normal distributions

The normal distribution is a mathematical function that is a particularly specified form of this phenomenon using two parameters:

- Mean (μ)
- Standard Deviation (σ)

It is defined by the following probability density function:

$$p(X|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} exp[-\frac{(X-\mu)^2}{2\sigma^2}]$$

$$X \sim N(\mu,\sigma)$$

Don't need to know this formula!

The normal curve is a defined distribution of values that has several mathematically useful properties.

It is NOT the same thing and is entirely distinct conceptually from "bell curves" in the wild!!!

Normal distributions

A normal distribution is a theoretical mathematical distribution with the following characteristics:

Central Tendency:

- The mean, median, and mode are all the same.

Variability:

- 68% of observations are within 1 SD
- 95% of observations are within 2 SDs
- 99% of observations are within 3 SDs

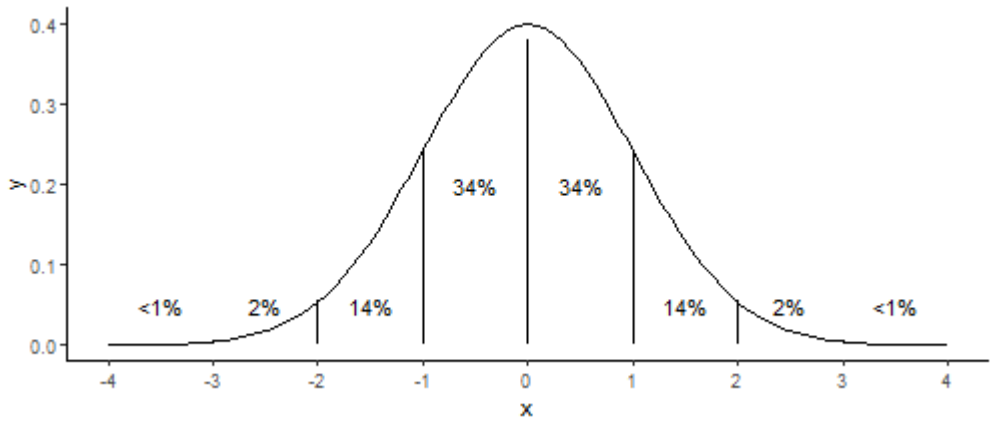
Shape:

- Unimodal
- Symmetrical
- Skewness & Kurtosis = 0

Normal distributions

More characteristics of the normal distribution:

- All members of a population fall within the normal distribution (it represents the whole of observations)
- Any area under the normal distribution curve corresponds to a defined percentage of observations that fall within that area.
- We can always infer the percentage of observations in any part of a normal distribution.



Empirical Rule

We can use our understanding of the normal distribution to calculate the probability of observing a value in a certain range. This is known as the empirical rule.

A randomly selected observation has an approximately...

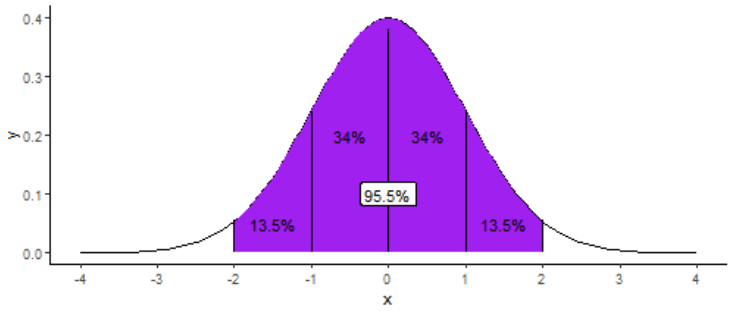
- **68% chance of being within 1 SD** of the mean.
  - $P(-1 < Z < 1) = 0.683$

Empirical Rule

We can use our understanding of the normal distribution to calculate the probability of observing a value in a certain range. This is known as the empirical rule.

A randomly selected observation has an approximately...

- **68% chance of being within 1 SD** of the mean.
  - $P(-1 < Z < 1) = 0.683$
- **95.5% chance of being within 2 SD** of the mean.
  - $P(-2 < Z < 2) = 0.955$

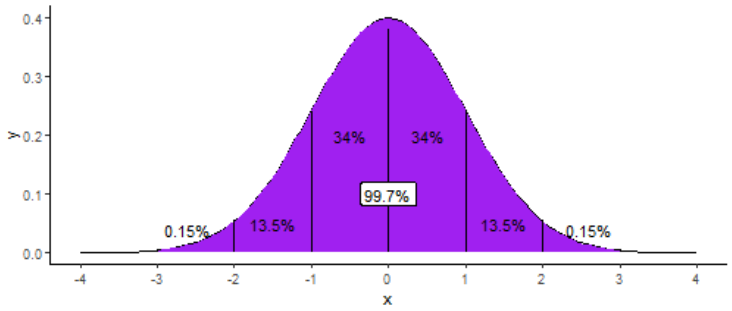


Empirical Rule

We can use our understanding of the normal distribution to calculate the probability of observing a value in a certain range. This is known as the empirical rule.

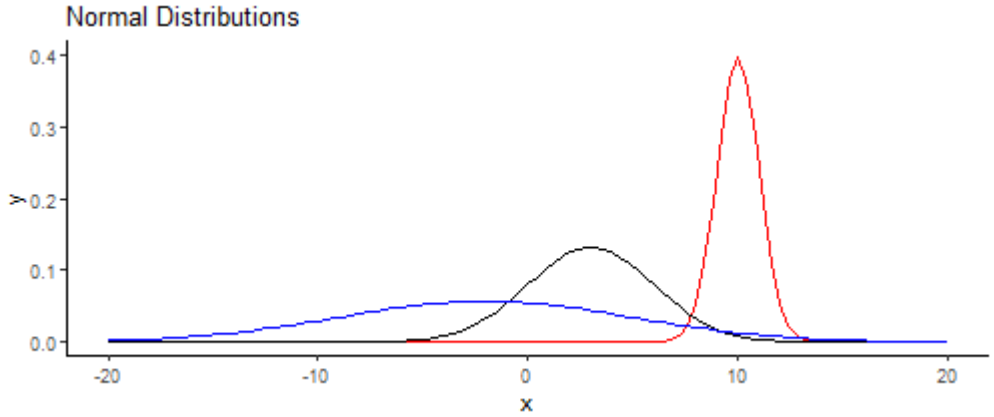
A randomly selected observation has an approximately...

- **68% chance of being within 1 SD** of the mean.
  - $P(-1 < Z < 1) = 0.683$
- **95.5% chance of being within 2 SD** of the mean.
  - $P(-2 < Z < 2) = 0.955$
- **99.7% chance of being within 3 SD** of the mean.
  - $P(-3 < Z < 3) = 0.997$



What is "normal"?

- A normal distribution can take on any mean and standard deviation.



- Regardless of the mean and standard deviation, the probability of selecting any particular value of that population at random always sums to 1.
- The empirical rule still applies (e.g., 68% of observations occur within 1 *SD* of the mean, 95.5% within 2 *SD* of mean, etc.).

Standard normal distribution

- The standard normal distribution has a mean of 0 and a standard deviation of 1.

- In this case, the x-axis represents z-scores, meaning an observation's value represents how far it is from the mean
    - A z-score of -1 indicates that an observation is 1 SD below the mean
    - A z-score of +2 indicates that an observation is 2 SD above the mean
- With a standard normal distribution, we can easily calculate the percentage of observations under any region

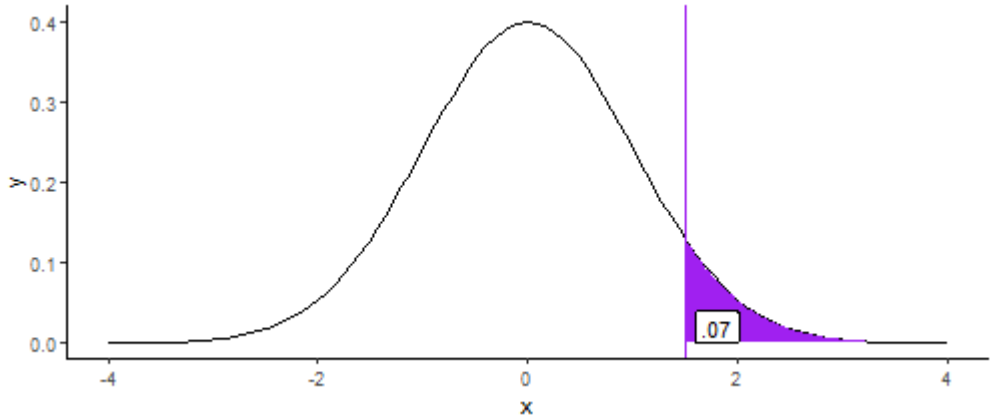
Area under the normal curve

Suppose we wanted to know the probability of scoring at least 1.5 SD above the mean on a standardized test of academic achievement.

To answer this question, we just need to find the area of the shaded portion under the curve.

Area under the normal curve

- We can either use some calculus or (more straightforward) a z-score look-up table to calculate the area under the normal curve.



Either approach would show the area above 1.5 SD of a standard normal curve is .07.

If the results of (say) a basketball skills test fell in a standard normal distribution, then this would imply that 7 percent of all individuals score 1.5 SD or higher on this test.

Given these few facts about normal distributions, we are now ready to move on to understand how and why **statistical inference** is possible...! **But first...**

You try

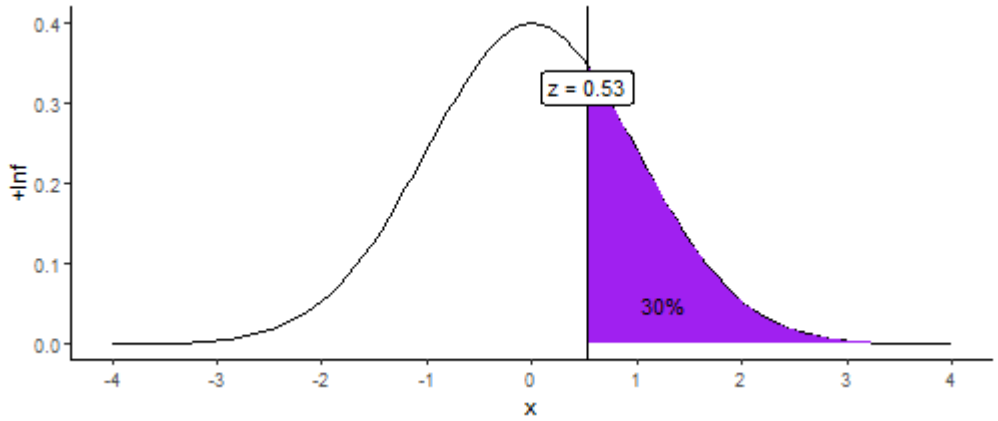
Let's try it with our WHO data:

- Suppose population mean national life expectancy is 71.64 (μ) with a SD of 8.15 (σ).
- **Assuming that country-level average life expectancies follow a normal distribution (spoiler: they don't!), what is the probability of randomly selecting a country with a life expectancy greater than 76?**
- First, we need to standardize our life expectancy value and give it a z-score. Then, use the z-score look-up table to answer the question.

$$z = \frac{76 - 71.64}{8.15} = 0.535$$

You try...answered!

The proportion of the area under the curve above a z-score of 0.53 is just under 0.30 ($\approx$ 0.296). Therefore, there is a 30 percent probability of randomly selecting a country with a national life expectancy greater than 76.



> "Ok, but this seems only helpful in the unrealistic case of observing a variable with a perfectly normal distribution. How does this help us in the real world?"

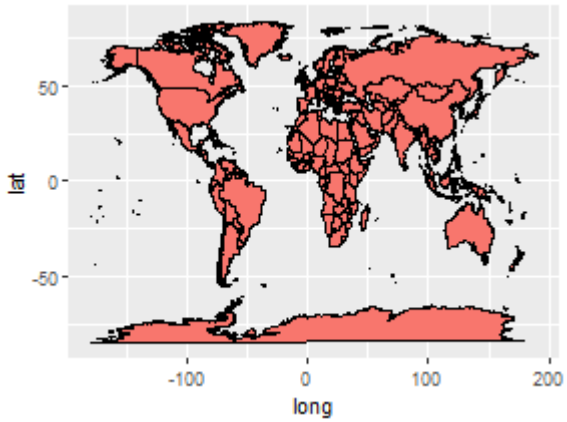
Get ready to have your mind blown by the beauty of the **Central Limit Theorem**!

Central Limit Theorem and statistical inference

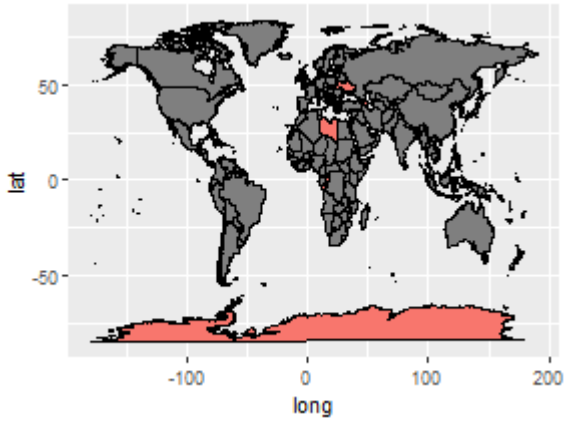
Population v. sample distributions

Using the WHO dataset, let's say we want to identify the mean life expectancy across all countries. However, let's also assume we only have the resources to sample from 10 countries.

Our **population statistic** represents the true value across all countries.[1]



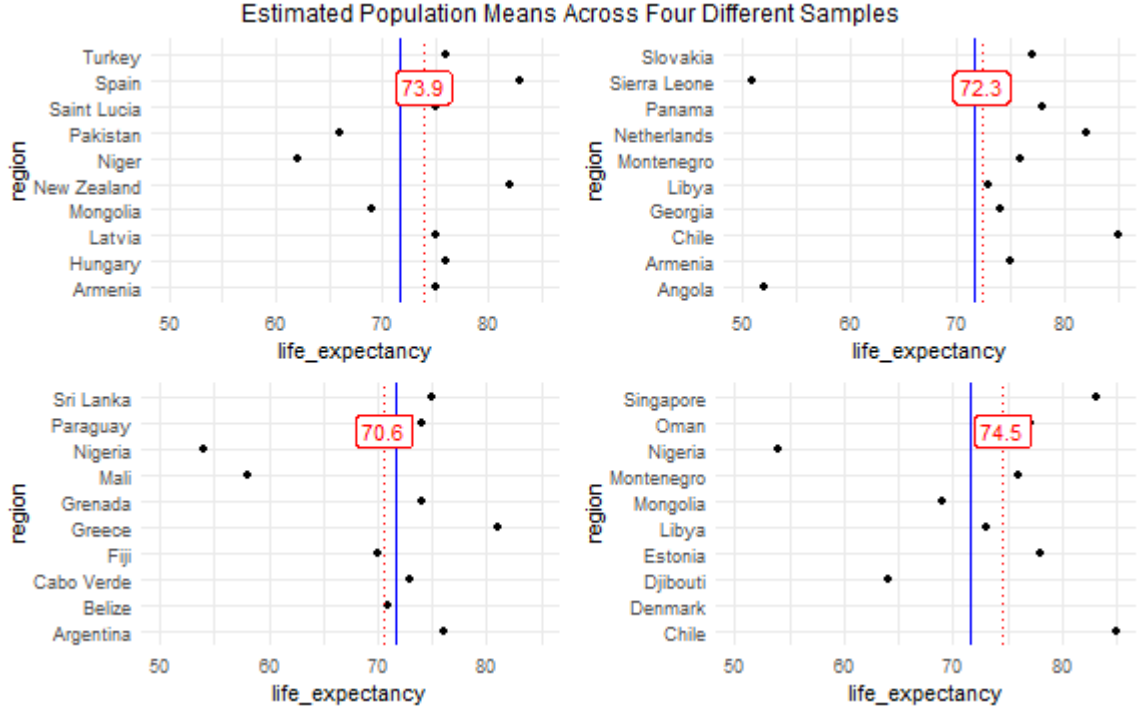
Our **sample statistic** will be our estimate of the population statistic using the countries we examine.



[1] Again, this is somewhat of a simplification...we're getting close to why...!

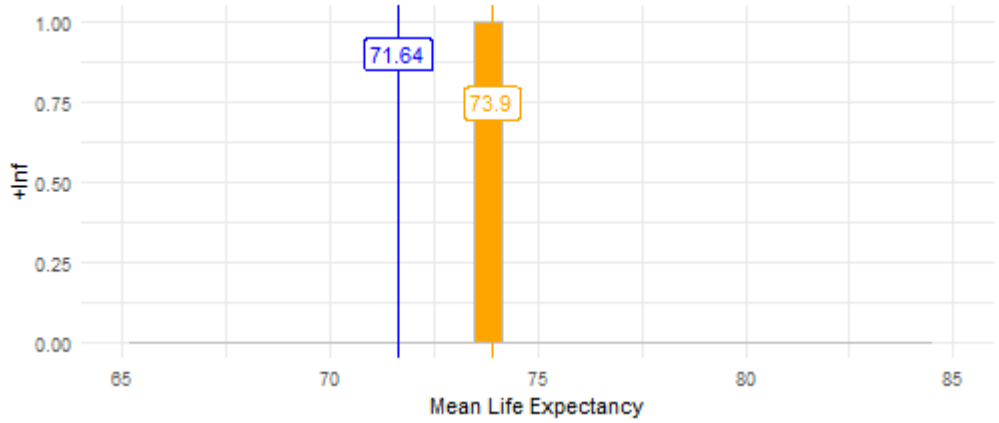
Sample estimates

- Assume that our true population mean (μ) is 71.64.
- Each possible random sample of 10 countries will produce its own sample mean ($\bar{x}$) or population mean estimate ($\hat{\mu}$).
- These estimated means have their own variability around the true population mean.



Population v. sample distribution

If we sample our population **100 times** and plot the estimated mean of each sample, the estimated means begin to form their own normal distribution. The mean of sample means is our **estimated population mean** $(\hat{\mu})$.



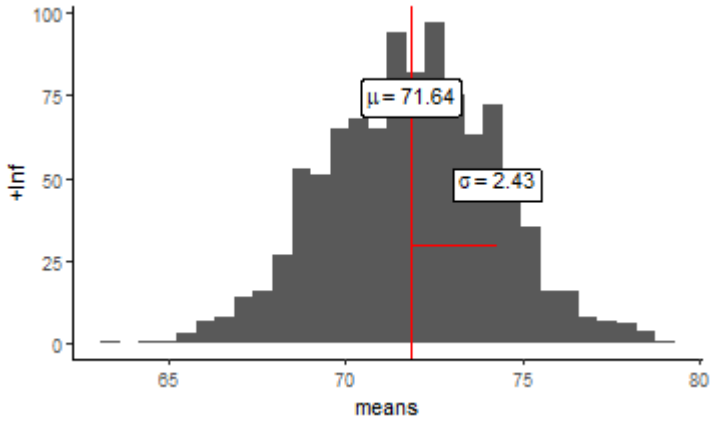
The mean of our sampling distribution (mean of means) approaches the **population mean** (μ).

If we sampled all possible samples of size *n* from our population, our mean of sample means would be equal to the population mean.

Central Limit Theorem (CLT)

- Random sampling from the population **will return** means that will be asymptotically (approaching) normal in their distribution as the number of samples approaches infinity. With a little bit of math, the Central Limit Theorem proves this fact.
- Because of that mathematical fact, we can conduct inference in statistics.
- We won't derive this formally, but instead will "prove" it by simulation (a few ways)

- Here is a histogram of the means of 1000 random samples from our WHO data.
  - The center of the distribution is our population mean.
  - The standard deviation is the standard error of the mean (SEM).



Even though our life expectancy distribution is only approximately normal, the distribution of our sample means **is** (approaching) normal.

CLT demonstration

See the following the link for more demonstrations of the central limit theorem:
https://onlinestatbook.com/stat_sim/sampling_dist/

Central Limit Theorem

> Given a population mean μ and standard deviation of σ, the sampling distribution of the mean is a normal distribution with a mean equal to μ and standard deviation equal to $\frac{s}{\sqrt{n}}$, where *s* is the standard deviation and *n* is the sample size.

- The distribution of sample means will approach the normal distribution as sample size increases regardless of shape of the population distribution. The standard error of the mean (SD of estimated means) will shrink as sample size increases.
- Center: Mean of sample means is the population mean: $\bar{X}_{\bar{x}} = \mu$
- Spread: Standard deviation of sample means is the standard error of the mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- The sampling distribution of means approaches a normal distribution because it is estimating a statistic of the population distribution. It does not "recreate" the shape of the population distribution.

Again, the observation that many populations take on "bell curve" shapes is different from the fact that the means calculated from repeatedly drawn samples form a normal distribution.

You interpret

> At your table, take turns explaining what you understand the Central Limit Theorem to show and why might this be useful for making statistical inference.

One sample z- and t-tests

z-test

A z-test is a statistical test to determine whether two population means are different when the variances are known and the sample size is large.

Let's try this with our WHO data:

> Suppose that we want to test whether the population mean life expectancy is greater than 70, and we know that the "true" population standard deviation is 8 years. We can draw a random sample of observations and pose the following null hypothesis:
>
> H_0: In the population, the mean life expectancy is 70.
>
> H_A: In the population, the mean life expectancy is greater than 70.

We draw a sample and calculate the average life expectancy in the sample: 71.64. We set our α-threshold at 0.05. **What does this threshold mean?**

We construct our z-score as follows:

$$\frac{71.64 - 70}{8} = 0.205$$

z-test

> Suppose that we want to test whether the population mean life expectancy is greater than 70, and we know that the "true" population standard deviation is 8 years. We can draw a random sample of observations and pose the following null hypothesis:
>
> H_0: In the population, the mean life expectancy is 70.
>
> H_A: In the population, the mean life expectancy is greater than 70.

Our z-score is 0.205.

How can we use our z-statistic look-up table to determine whether there is a less than 5 percent chance that our observed mean life expectancy is greater than 70?

What do you conclude from your test?[1]

[1] *We are going to walk through the steps in class. Just looking at the slides is not sufficient for you to understand how to do this.*

Population variance

- In a z-test, the population standard deviation is assumed to be known.
- In practice, we **rarely** know the true population standard deviation **even if we observe all individuals in our "population"** (e.g., all students in a state, all countries in the world). This is because:
    - Measurement error (e.g., we didn't measure how long everyone lived accurately)
    - Population statistics can be in constant flux (e.g., average life expectancies can change moment to moment)
    - There are elements of idiosyncrasy in each individual's realized outcome that might have come out differently in some other lived reality (e.g. Sliding Doors, multi-verses)
    - Can be unfeasible to measure entire population

Thus, we tend to use **Student's t-distributions** to make our inferences, because these do not require us to know population variability.[1]

[1] The name comes from a pseudonym ("Student") used in papers written by William Sealy Gossett, the Head Experimental Brewer at Guinness, in which he developed the statistical theory around the t-distribution. His interest in statistics was in improvements in the cultivation of barley grains. I have reviewed several biographical articles about him and none attribute eugenic beliefs to him, but he was friendly with Fisher and Pearson, so proceed with caution.

t-distributions

- t-distributions assume that the *population* standard deviation (σ) is unknown. Instead, it uses the *sample* standard deviation (s).
- t-distributions with smaller sample sizes have "fatter" tails. Assumes more uncertainty with smaller sample sizes.
- t-distributions resemble z-distributions as degrees of freedom increase.

Degrees of freedom

The degrees of freedom refers to the number of values in the calculation of a statistic that are free to vary.

In the calculation of simple, one-sample statistics (e.g., the mean) our degrees of freedom are always $n - 1$, where n is our sample size.

This is because once you know the mean and all but one of the values of the observations, the last observation must be a defined value:

$$mean(71, 73, 76, 78, ??) = 75.4$$

$$\frac{71 + 73 + 76 + 78 + ??}{5} = 75.4$$

$$?? = (75.4 * 5) - 71 - 73 - 76 - 78$$

$$?? = 79$$

In this case, therefore, our degrees of freedom are $5 - 1$; i.e. 4.

Unbiased variance estimates

As you'll recall, the formula for the population standard deviation we used before was:

$$\sigma = \sqrt{\frac{\Sigma_{i=1}^{n}(x_i - \mu)^2}{N}}$$

but, when we calculate this in reality, we are using the sample mean $(\bar{x})$ which is constructed **using each** x_i!

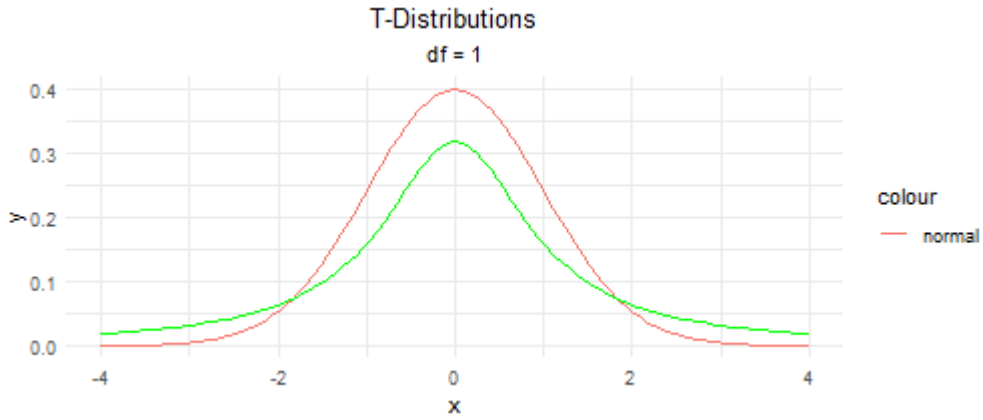
You can read a more detailed statistical treatment of Bessel's correction, but the key takeaway is that using sample statistics generates bias (inaccuracy) in the estimation of the population statistics.

To produce an unbiased estimate of the mean, variance or standard deviation, we should always subtract one from the sample size in the denominator. For example, when we construct the standard deviation of a sample, we should calculate it as:

$$s = \sqrt{\frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

t-distributions

- Likewise, the degrees of freedom for a t-test is always *n*–1, where *n* is our sample size.
- t-distributions with fewer degrees of freedom have "fatter" tails.



As n increases, the t-distribution approaches (asymptotically) the normal distribution. In large-ish samples (basically >650), the t- and z-statistics and associated tests are essentially indistinguishable.

One sample t-test

- t-tests are one of the most foundational statistical tests in classical statistics.
- The purpose of a one-sample t-test is to describe the probability of obtaining a particular sample mean through random sampling variability, assuming the population mean = K.
    - e.g. Is our observed sample mean to be expected if we repeatedly drew random samples from the population?

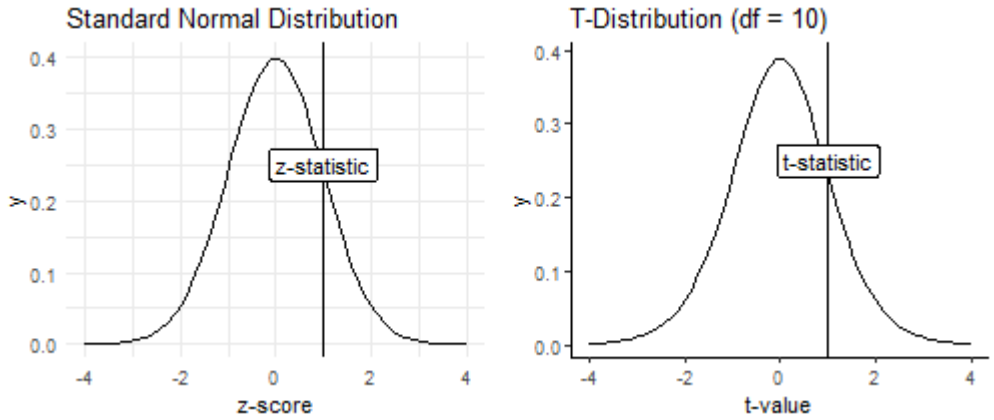
Example Research Question: Suppose the known average life expectancy is 76 years. We collect a sample of all countries' life expectancy, and these have a standard deviation of 3.2 years. We would like to know whether countries around the Mediterranean Sea significantly differ from the known population average of 76?

One sample t-test

- The one sample t-test does **not** test the probability of obtaining the sample distribution or data itself, only the sample mean.
- Using the probability of obtaining a particular sample mean, our **Null Hypothesis** (H_0) is $\mu = K$, where K is any constant.

One sample t-test

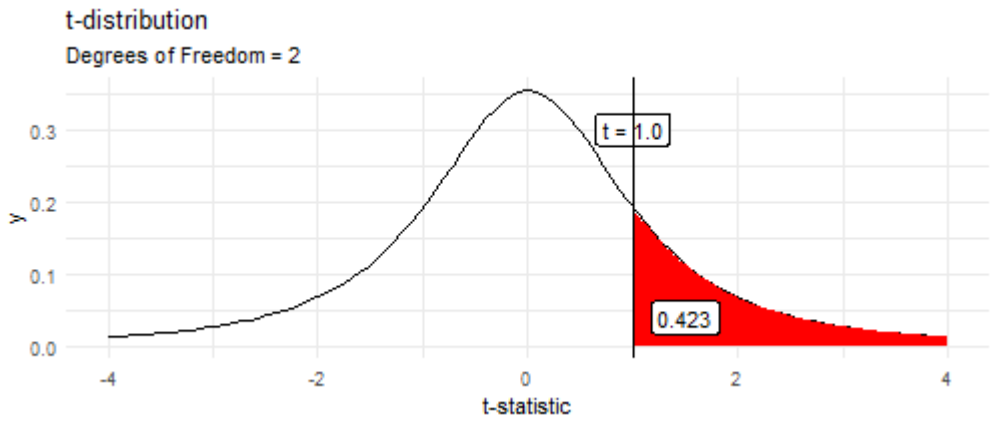
- With a z-test, we were able to find the number of observations that fell beyond a particular value. That value was our z-statistic.
- The t-statistic is functionally similar to the z-statistic, but for the t-distribution.



t-distribution

- Shape of the t-distribution is slightly different depending on the degrees of freedom $(n-1)$.
- Thus, the proportion of the area under the curve beyond our t-value also depends on our degrees of freedom.

Here is a similar animated graphic as on slide 34, but now highlighting the changing proportion of the curve falling beyond a given t-statistic; in this case $t = 1$.



You can look up the critical t-value for any number of degrees of freedom and alpha-threshold in any number of t-statistic tables you can find online.

One sample t-test

To calculate a t-statistic, we use the same formula to calculate a z-score, but replace the population standard deviation with the sample's standard deviation.

Since we don't know the population standard deviation (σ), we use our sample's standard deviation (s) as an estimate of the population standard deviation $(\hat{\sigma})$. The larger our sample (and more degrees of freedom we have), the less we need to worry about mismeasurement of our sample standard deviation.

Thus, this is our t-statistic:

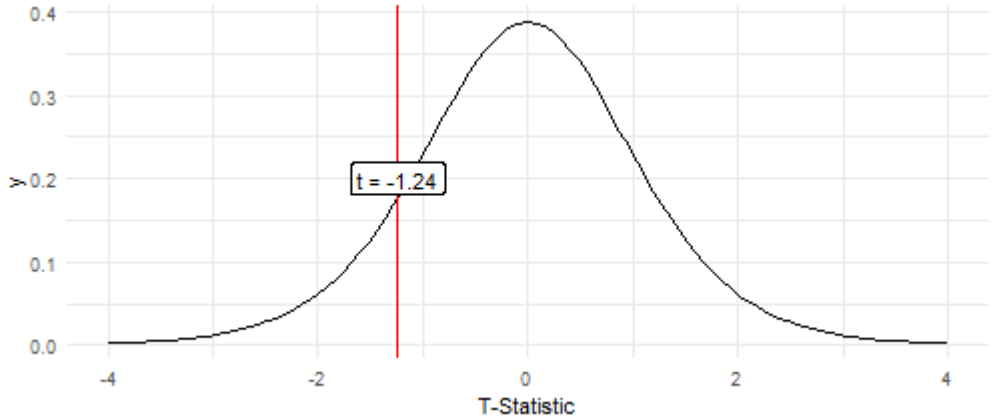
$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Applying a t–test

Suppose we have data from a sample of 10 countries around the Mediterranean with a mean life expectancy of 74.75 and standard deviation of 3.2.

If our population mean is thought to be 76 ($H_0 : \hat{\mu} = 76$), what is the probability of obtaining a sample mean of 74.75, **or a more extreme mean value**, due to random sampling variability?

$$t(9) = \frac{74.75 - 76}{3.2/\sqrt{10}} = -1.24$$



Applying a t-test

Assuming a population mean of 76 and sample size of 10...

...the proportion of random samples that would demonstrate a sample mean **equal to or lower than 74.75** is 0.12 (i.e., p = 0.12).

...the proportion of random samples that would demonstrate a sample mean **equal to or more extreme than 74.75** is 0.24 (i.e., p = 0.24).

Note: One-tailed tests are rarely warranted and we will generally be conducting two-tailed tests.

> If the null hypothesis is true, the probability of obtaining a sample mean equal to or more extreme than 74.75 is 0.24.

One-tailed vs. two-tailed test

A one vs. two tailed test apply different alternative hypotheses from the null.

$$H_0 : \mu = K$$

One-Tailed

- $H_A : \mu < K$ or $\mu > K$

Tests for a significant difference in one direction. Assumes *a priori* that the difference can only occur in one direction.

Two-Tailed

- $H_A : \mu \neq K$

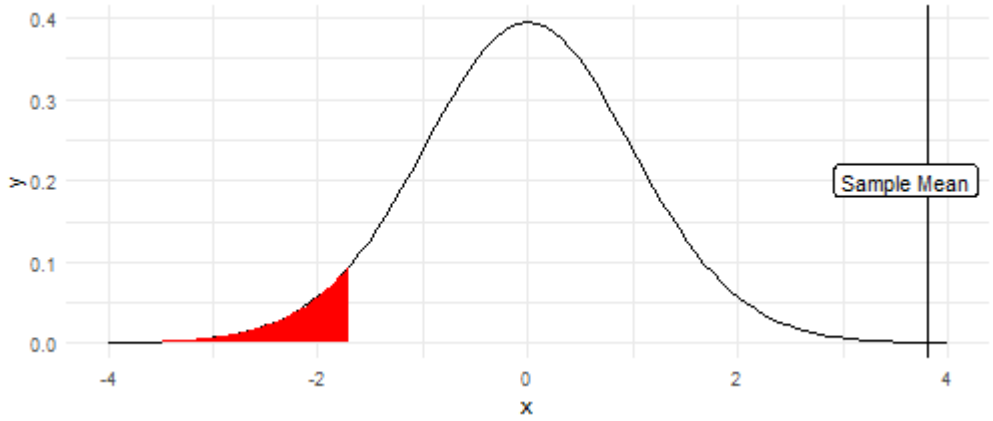
Tests the possibility of a significant difference in either direction.

Demonstration

- Dr. DSM and Dr. APA have developed a new intervention for depression.
- They would like to compare the post-treatment depression ratings for their research participants to average depression ratings following "business as usual."
- They conduct a one-tailed t-test to see if their sample's mean post-treatment rating is significantly better than the normal post-treatment averages.
    - $(H_0 : \mu < K)$
    - Lower scores depression ratings are desirable.

Demonstration

Dr. DSM finds that the mean post-intervention depression rating did not fall below the critical t-value for the one-tailed test. The new intervention group did not demonstrate significantly lower than normal post-treatment averages.

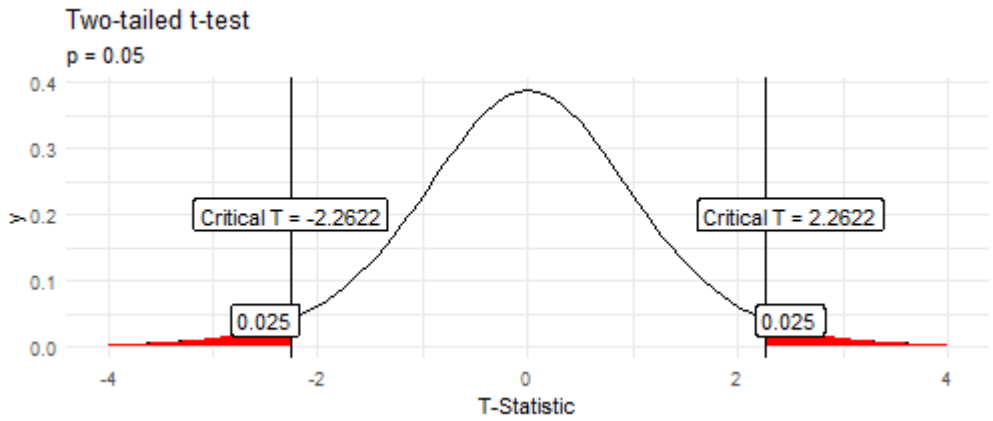


What conclusion was missed with the one-tailed test? The sample had much higher levels of depression than we'd expect! Neither the null ($\mu = K$) nor the alternative hypothesis ($\mu < K$) seem tenable.

Alpha thresholds and t-values

With two-tailed tests, the total alpha is "split" across both tails of the distribution.

- If $\alpha = .05$, $\alpha/2 = 0.025$.
- Critical t-values $t_{\alpha/2}(df)$ are the values at which our alpha threshold is met.

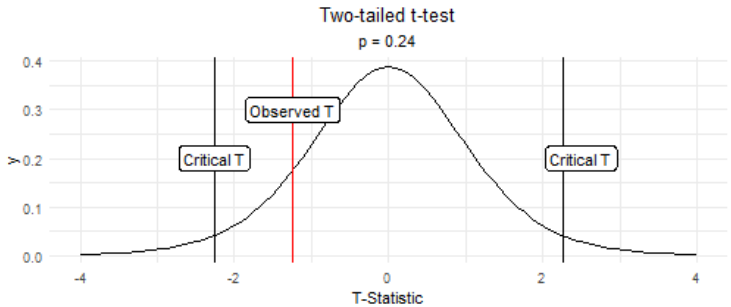


If we conducted a t-test, our t-statistic would have to be less than -2.2622 or greater than 2.2622 to reject the null hypothesis.

One sample t-test

- Assuming we had set our α at .05, let's see if we reject our null hypothesis in our earlier question: *do countries around the Mediterranean Sea significantly differ from the known population average?*

Our observed t-value did not surpass our critical t-values.



Our p-value is not less than .05 and does not satisfy our alpha threshold.

Whether we compare our t-statistic to our critical t-values or look at our p-value, we come to the same conclusion.

Interpretation of t-test

Assuming a population mean of 76 and sample size of 10, the proportion of random samples that would demonstrate a sample mean **equal to or more extreme than 74.75** is 0.24 $(t(9) = -1.24, p = 0.24)$. Because our p-value does not meet our alpha threshold of .05, we **fail to reject the null hypothesis**.

Thus, we can conclude that the mean life expectancy of this sample of countries around the Mediterranean is not different from the mean population average.

Implementing in R

Let's examine in our data whether the sample of countries labelled as economically "Developing" have a different average life expectancy than the "true" known population average.

```
low_inc <- filter(who, status=="Developing")

t.test(low_inc$life_expectancy, mu = 71.64)
```

```
##
##     One Sample t-test
##
## data:  low_inc$life_expectancy
## t = -3.1667, df = 150, p-value = 0.001868
## alternative hypothesis: true mean is not equal to 71.64
## 95 percent confidence interval:
##  68.49272 70.91125
## sample estimates:
## mean of x
##  69.70199
```

Implementing in R

Let's examine in our data whether the sample of countries labelled as economically "Developing" have a different average life expectancy than the "true" known population average.

```
# The defaults for t.test are to assume that
# you are conducting a two-sided one-sample test,
# with an alpha threshold of 0.05. You can modify as needed:

t.test(low_inc$life_expectancy, mu = 71.64, alternative = "less")
```

```
##
##     One Sample t-test
##
## data:  low_inc$life_expectancy
## t = -3.1667, df = 150, p-value = 0.0009341
## alternative hypothesis: true mean is less than 71.64
## 95 percent confidence interval:
##     -Inf 70.7149
## sample estimates:
## mean of x
##  69.70199
```

Putting it together

1. We can characterize the central tendency and variability of any distribution with several useful statistics
2. The normal distribution has several features that allow us to state precisely how likely it is that we would observe any particular value under the normal curve
3. Statistics calculated in samples that are drawn repeatedly from a population will form their own distributions.
4. Even if the population itself is not normal, the mean of sample means drawn from that population **will** be normal
5. As a result, we can test whether any observed sample mean is different from the population mean
6. Depending on whether or not we known the population variance, we can test this using either a z- or a t-test statistic

Even if it's not apparent yet, these foundational building blocks will allow you later to compare two groups, characterize the relationship between two variables and more!

Synthesis and wrap-up

Class goals

- Describe special features of a normal (and standard normal) distribution
- Interpret a z-statistic table
- Describe the distribution of repeated sample statistics drawn from a population, how this relates to the Central Limit Theorem (CLT) and how this is informative to statistical hypothesis testing
- Determine whether the mean value of a sample is different than a defined population mean, both when the population standard deviation of the variable is known (z-test) and when it is unknown (one-sample t-test)

To Dos

Reading

- LSWR Chapter 10: inference and CLT

Quiz

- Quiz #3: Opens 3:45pm on Oct. 31, closes at 5pm Nov. 1
- QuiZ #4: Opens 3:45pm on Nov. 14, closes at 5pm Nov. 15

Assignment

- Assignment #3 Due Nov. 9, 11:59pm