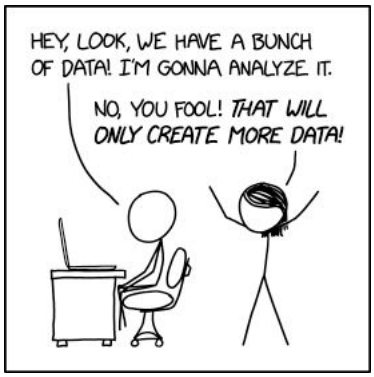
Introduction Statistics Crash Course

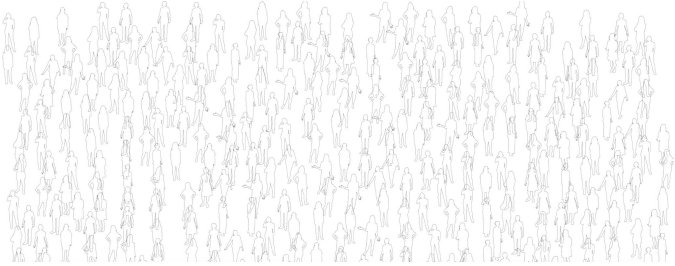
What we need to get a sense of ahead of the workshop:

1. Basic premise of statistics.
2. Our favorite probability distribution
3. Sampling distribution
4. Confidence intervals
5. Hypothesis tests
6. Test statistic
7. Our other favorite probability distribution



What is the big picture when it comes to statistics?

Population - The entire collection of individuals that are of interest.



Sample - The subset of the population for which data can actually be obtained.

If we want the sample to be representative (or “look like”) the population of interest, the sample should be random.

What is the big picture when it comes to statistics?

Parameter - a number summarizing some feature of a population (like an average or a proportion) as if we observed the feature for every member of the population

Statistic - the corresponding number for the sample, i.e., the summary based on what we observe in a sample, it is our best "estimator" for what the value of the parameter is

If the sample is representative of the population we expect the statistic to be a good estimate of the parameter.

The distribution of our sample data (what we see) is an approximation of the distribution of the population data (what we do not know).

Based on prior knowledge we may have an idea of the shape of the population distribution.

As our sample gets larger, we get a more accurate representation of what the population distribution is.

The distribution of the population is often described by a density curve.

Density Curves

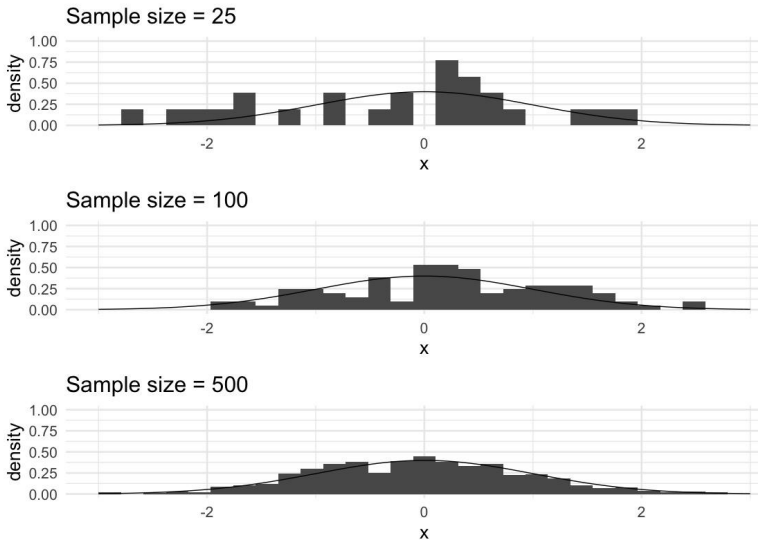
A density curve is a mathematical model (equation) that describes a distribution.

To make a density curve, we rescale a percent histogram so that the total area under the curve is 1.

The area under the curve between a range of values is the proportion of all observations that fall in that range.

A density curve must:

- have area under its curve equal to 1, and
- always lie above the x-axis.



Proportions and probabilities

The area under the curve between a range of values is the proportion of all observations that fall in that range.

We will use the following interchangeably:

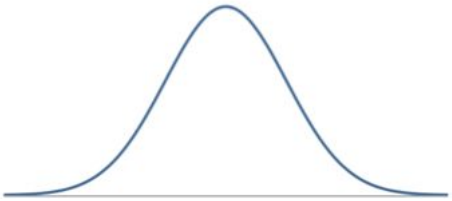
- What proportion of students are shorter than 61 inches?
- What is the probability that a randomly selected student is shorter than 61 inches?

The probability of any outcome of a random phenomenon is the proportion that an outcome occurs in a long series of repetitions.

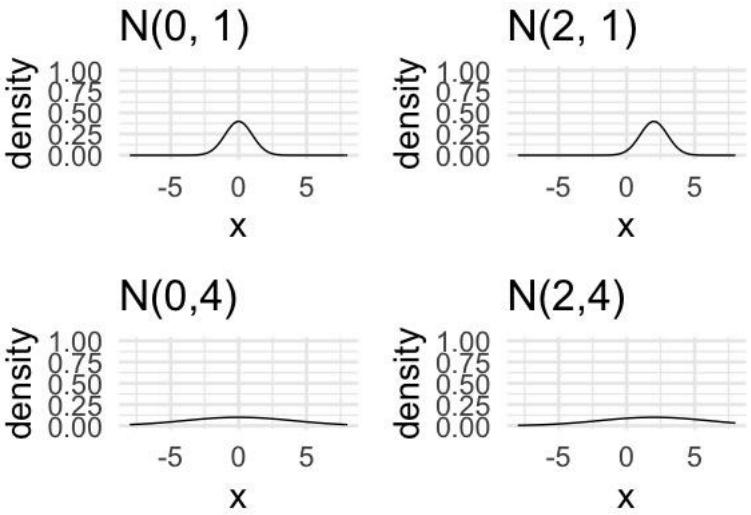
Note: $P(X < 2) = P(X \leq 2)$

The Normal Distribution (aka Gaussian distribution or bell curve)

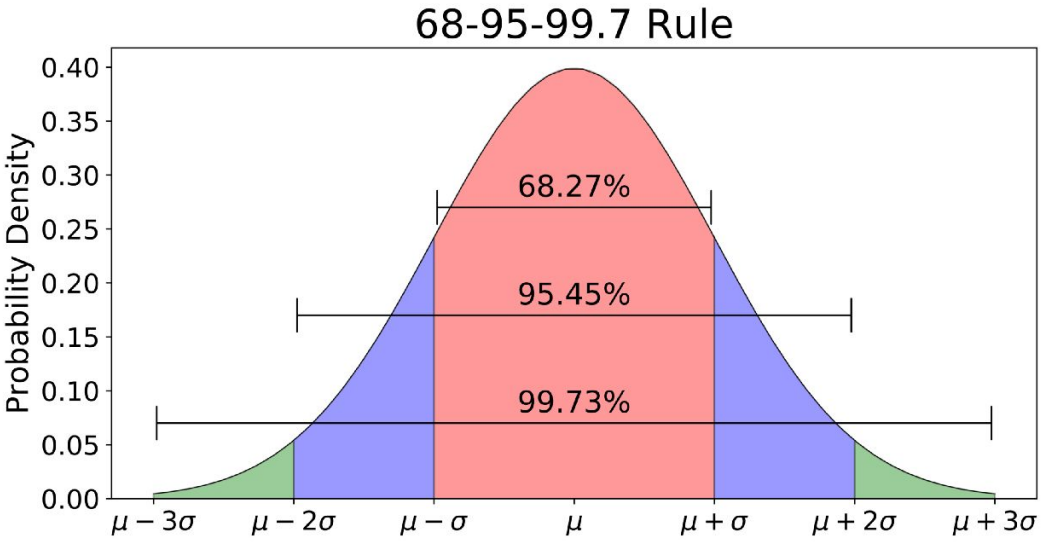
- Unimodal and symmetric, bell shaped curve
- Many variables are nearly normal, but none are exactly normal
- Defined by two parameters and denoted as $N(\mu, \sigma)$ → Normal with mean μ and standard deviation σ



The normal distribution is a family of curves.

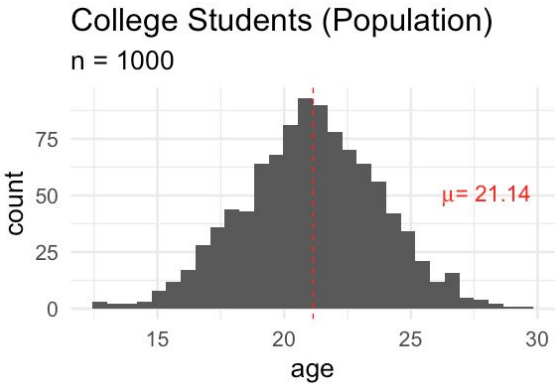


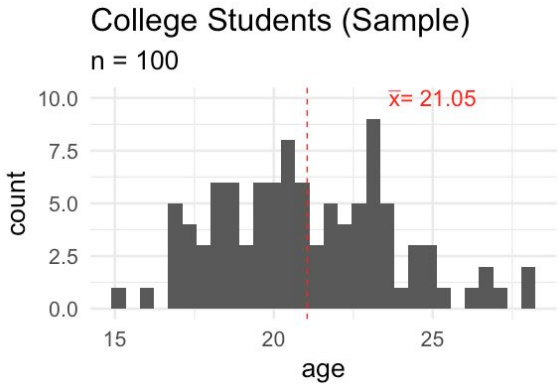
All normal curves follow this rule, telling us about the area under very specific parts of the curve.



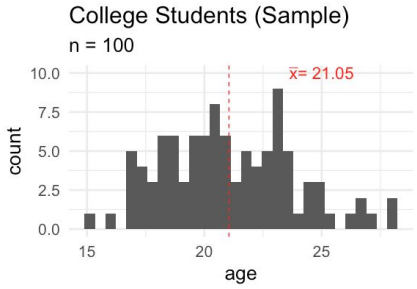
Sampling distribution - the distribution of values taken by a _____ in all possible samples of the same size from the population

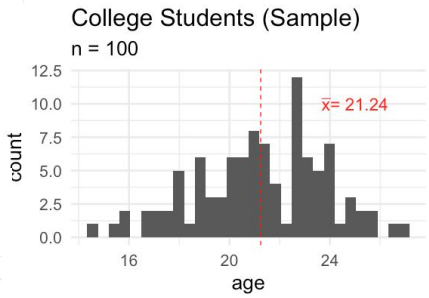
How far away from the truth is our estimate?

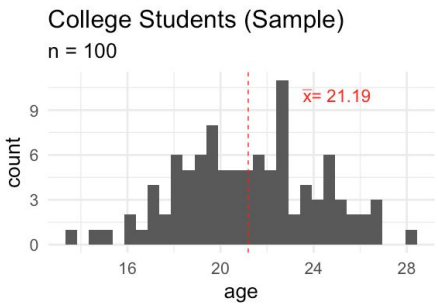




How does x-bar vary across different samples from the population?

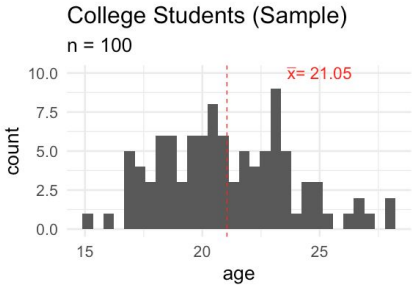


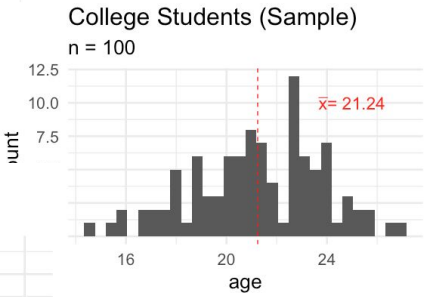


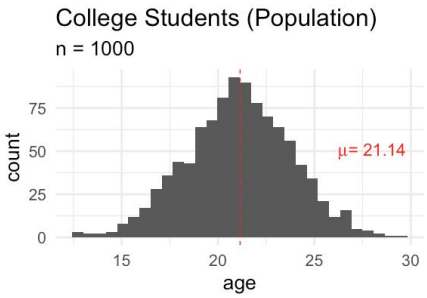


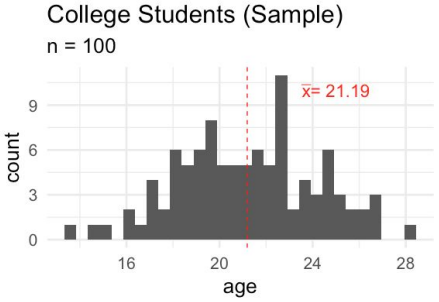
How does x-bar vary across different samples from the population?

How far from mu do we expect x-bar to be based on the variability induced by the sampling?









What shape is the sampling distribution? The Central Limit Theorem (CLT)

When sampling randomly from _____ with mean mu and standard deviation sigma, if n is large enough, then the sampling distribution of x-bar is approximately normal.

How large is large "enough?

It depends! But in general _____ observations are required if the population distribution is far from normal.

A sample of size _____ is generally enough to obtain a normal sampling distribution from strong skewness or mild outliers.

A sample of size _____ will typically be good enough to overcome extreme skewness and outliers.

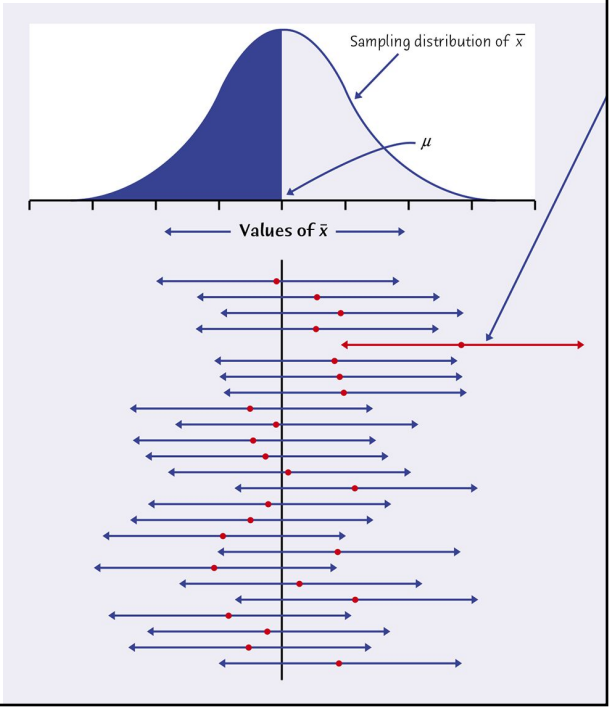
We can say that the population mean mu will be within 1.96 standard deviations from the sample mean x-bar in approximately 95% of all samples.

- The population mean mu will fall in the interval of approximately 95% of all samples.
- Hence the population mean mu will fall outside the interval in approximately 5% of all samples

Therefore, for the interval built on x-bar from our single sample, we can be 95% confident that the interval contains the true mean mu.

This is what allows us to perform statistical inference.

The suggested interval only depends on known quantities ( ), not the unknown mean mu.



Language matters - how to read a confidence interval

We are ____ confident that the population parameter (_____) is between _____.

Informally: a range of plausible values for the population parameter

What are we NOT saying?

Not a statement of the probability that the truth lies in our interval. It either does or it doesn't. We don't know which case our particular interval is, hence the uncertainty language.

Only a statement about capturing the population parameter, not an individual observation or proportion of observations.

A test of statistical significance (hypothesis test) tests a specific hypothesis about a population parameter using the sample data to decide on the validity of the hypothesis.

Think about a court of law:

- Competing “hypotheses”
- Believed innocent until proven guilty
- To be found guilty: proof beyond a reasonable doubt
- Verdict:

A test of statistical significance (hypothesis test) tests a specific hypothesis about a population parameter using the sample data to decide on the validity of the hypothesis.

Proof by contradiction

- Assume the claim about the population parameter is true (hypothesis is true).
- Show that it would be very unlikely to have obtained your sample, if that assumption were true.
- Conclude that the original claim must not be true.

Statistical Hypotheses

In significance testing, you test one hypothesis versus another. They are competing beliefs.

Null Hypothesis

- The status quo, or what is currently believed: a statement of “nothing is happening”, “no effect”, “no difference”.
- We look to disprove or reject this hypothesis.
- Labeled (H-zero or H-naught)
- The null hypothesis always includes an equals sign.

Alternative Hypothesis

- The claim we are trying to find evidence for; a statement of “something is happening”.
- We look to reject the null hypothesis in favor of the alternative.
- Labeled (H-a)
- It can be a one-sided (< , >) or two sided (not equal) hypothesis.

General Form of Hypotheses

A two-tailed or two-sided hypothesis test of the population mean has the hypotheses:

A one-tail or one-sided test of a population mean has the hypotheses:

- We use to denote a specific number - it is just a placeholder. We do not use it in any actual statement of hypotheses for a specific problem.
- Before you state your hypotheses, you must define the parameter mu, the true mean of the population in the context of the problem.

P-value: the probability that we would observe a sample mean this extreme or more extreme (in the direction of the alternative), if the null hypothesis were true, these are all 'tail' probabilities

Two-Sided

One-Sided

P-values and their interpretations

Small P-value

- With a small p-value, we _____.
- We have sufficient evidence in favor of the _____.
- Small p-values provide strong evidence _____ the null hypothesis and in support of the alternative.

Large P-value

- With a large p-value we _____.
- We have insufficient evidence to conclude that _____ holds.
- Does NOT give us evidence that the null is true.
- Suggests that random variation alone may account for the observed differences between the sample and population means. This is not the same as saying the null is true.
- Large p-values fail to provide evidence _____ the null hypothesis.

We never *accept* anything and conclusions are always about your decision regarding the null hypothesis.

Significance level - alpha

The significance level, is the largest p-value tolerated for rejecting a true null hypothesis (how much evidence is required against the null). This value is decided _____ conducting the test.

- If the p-value is less than or equal to alpha, then we _____.
- If the p-value is greater than alpha, then we _____.

When the shaded tail area (p-value) becomes very small, the probability of drawing such a sample by chance, given that the null is true, gets _____.

A p-value less than a significance level alpha is considered **statistically significant,** i.e. observed phenomenon is _____ to be entirely due to chance event from the random sampling.

Our one-sided test gave a p-value of 0.0228.

- If alpha had been set to 0.05, then the p-value would be significant.

- If alpha had been set to 0.01, then the p-value would *not* be significant.

General Forms of Inference

Confidence Interval

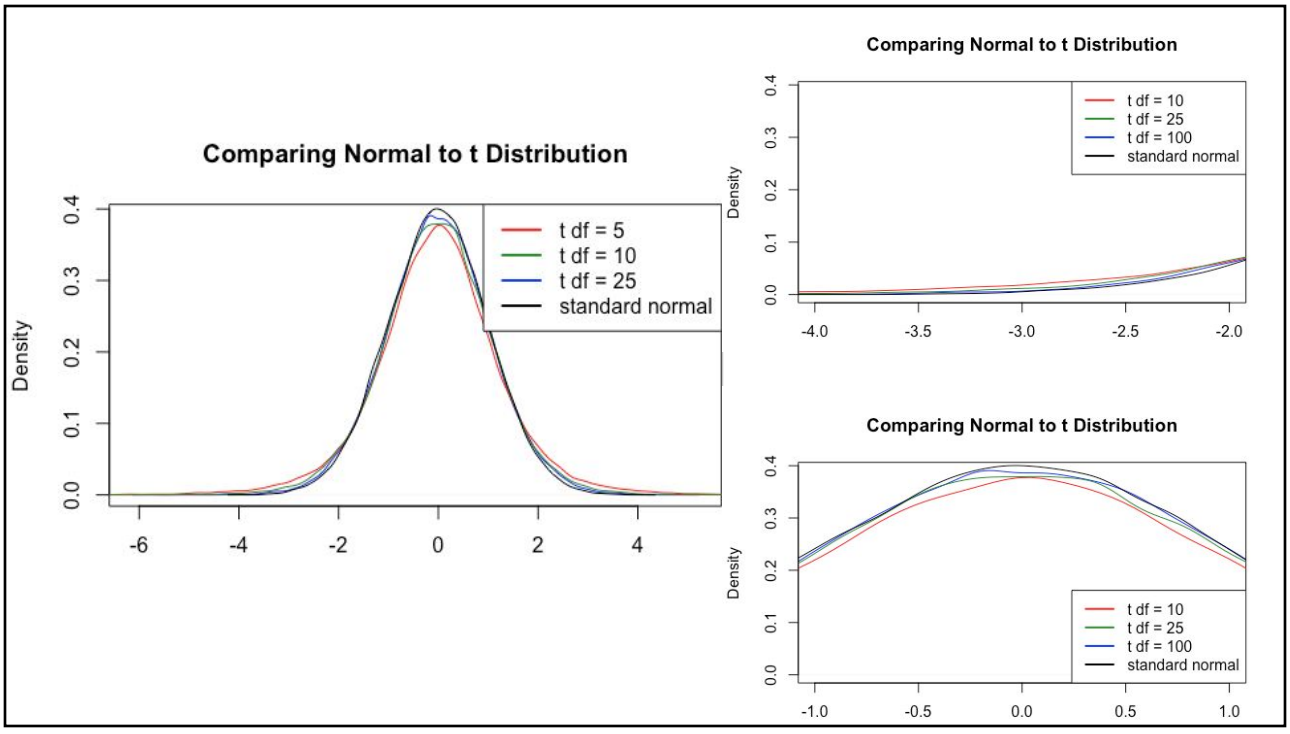
Test Statistic

When the population standard deviation is unknown...

- We use the sample standard deviation, s, to estimate the population standard deviation, sigma.
- The standard error of the sampling distribution of x-bar becomes _____.
- Replacing sigma with s adds some extra uncertainty to our inference (we are estimating two things instead of one).
- Because of that extra uncertainty, the standardization of x-bar no longer results in a normally distributed test statistic.

Properties of the t-distribution

- The t-distribution is a bell-shaped distribution centered at 0.
- The parameter for a t-distribution is called the _____ and this parameter essentially controls the variability in the distribution.
- The t-distribution is _____ than the Normal distribution.
- The t-distribution has _____; the degrees of freedom indicate how “heavy” they are.


Comparing Normal to t Distribution
Density
0.0
0.1
0.2
0.3
0.4
-6
-4
-2
0
2
4
t df = 5
t df = 10
t df = 25
standard normal
Comparing Normal to t Distribution
Density
0.0
0.1
0.2
0.3
0.4
-4.0
-3.5
-3.0
-2.5
-2.0
t df = 10
t df = 25
t df = 100
standard normal
Comparing Normal to t Distribution
Density
0.0
0.1
0.2
0.3
0.4
-1.0
-0.5
0.0
0.5
1.0
t df = 10
t df = 25
t df = 100
standard normal