Stat 230 Introductory Statistics Sampling Distributions, Central Limit Theorem, and Confidence Intervals

University of British Columbia Okanagan

Term 2, 2019S

Introduction to Sampling Distributions

- Suppose we want to work out the average height of men on the UBCO campus, do we measure all of the men on campus?
- It is often impractical to measure the entire group of interest
- Instead, we could consider a random sample that is *representative* of the population under consideration
- This sample is then used to make *inferences* about the population.

Population Distribution

Definition 5.1

The **population distribution** is the distribution of the realized values of a variable for all individuals of a population.

So in the example of height of men on the UBCO campus,

- the population is all the men on UBCO campus
- the variable of interest is their height
- the population distribution is the distribution of these heights.

Population Distribution vs. Sampling Distribution

- The **population mean**, denoted by Greek letter μ, is found by taking the arithmetic mean of values of the variable of interest for the *entire population*.
- Similarly, the **population standard deviation**, denoted by Greek letter σ, is found by calculating the standard deviation of the measurements taken on the *entire population*
- The **sample mean**, denoted $\overline{x}$, is found by taking the arithmetic mean of values of the variable in our *sample*.
- Similarly, the **sample standard deviation**, denoted s, is found by calculating the standard deviation of the measurements in our *sample*.

Sampling Distribution

- The members of our sample are taken at random, therefore our summary statistic (in this case the mean) will be random as well.
  - We don't expect $\overline{x}$ to be exactly equal to μ
  - Nor do we expect several samples to yield the exact same $\overline{x}$
- In other words, $\overline{X}$ is a random variable!
- Recall we use capital letters, e.g. $\overline{X}$, for random variables before they are observedand lower case for realizations of that random variable e.g. $\overline{x}$ (obtained after we take a sample and compute the sample mean)

Sampling from a Normal Distribution

The **sampling distribution** of the mean gives the distribution of $\overline{X}$

In other words, it gives the theoretical distribution of $\overline{x}$ had we taken many an infinite number of samples (producing an infinite number of different $\overline{x}$'s).

Theorem 5.2

If we take a random sample of size n from a normal distribution with mean μ and standard deviation σ then the distribution of the sample mean is

$$\overline{X} \sim N\left(\mu_{\overline{x}}, \sigma^2_{\overline{x}}\right),$$

where $\mu_{\overline{x}} = \mu$ and $\sigma^2_{\overline{x}} = \sigma^2/n$.

This can easily be seen from our rules of Expectation and Variance

...

Example 5.1

Suppose the height of men on campus is normally distributed with mean 1.78m and standard deviation 0.09m. What is the probability that the mean of a sample of 25 men will be < 1.8m?

Central Limit Theorem (CLT)

Definition 5.3

If we take a random sample of size n from **any** population with mean μ and standard deviation σ then when n is large, the distribution of the sample mean is approximately Normal with mean μ and variance σ^2/n. Hence for sufficiently large n,

$$\overline{X} \text{ is approximately } N(\mu, \sigma^2_{\overline{x}}),$$

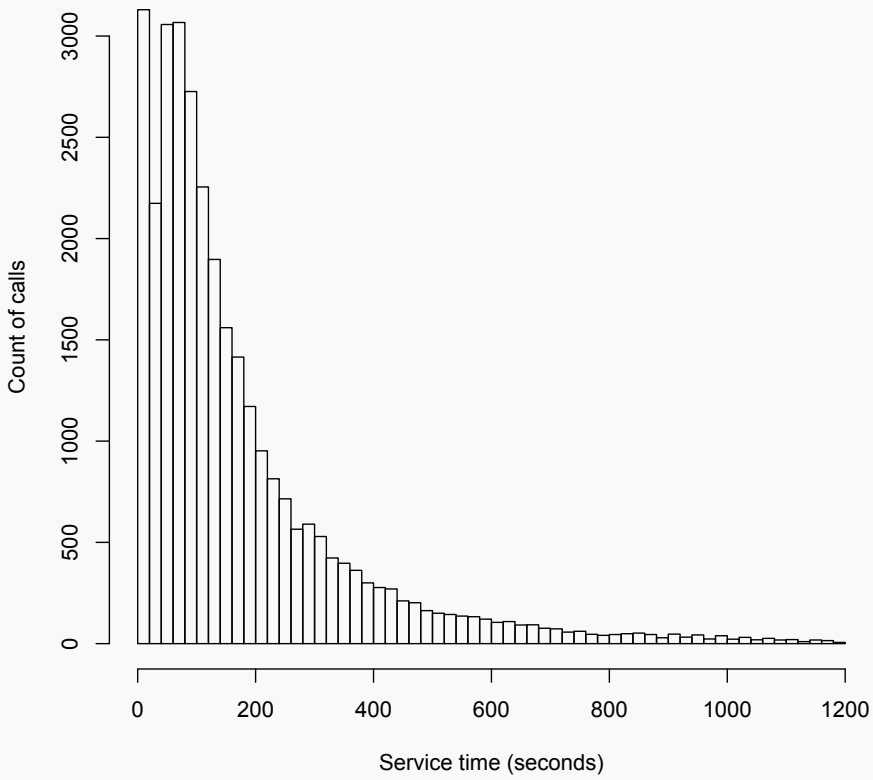
where $\sigma^2_{\overline{x}} = \sigma^2/n$.

- Note that this makes no assumption of the population distribution

Central Limit Theorem (CLT)

Let's see a demonstration. The histogram on the right plots the length in seconds of 31,248 calls to a bank's customer service centre.[1]

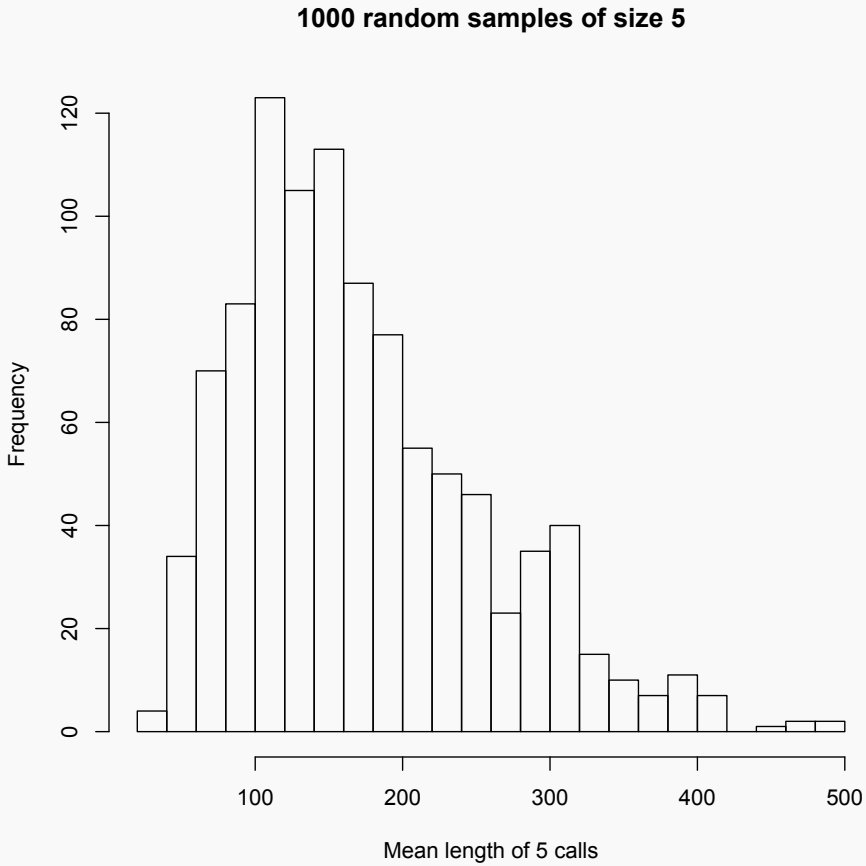
- The data are clearly skewed right
- Certainly not normal (bell shaped and symmetric).
- If we consider the 31,248 calls our 'population', then the population mean is $\mu = 173.9509$

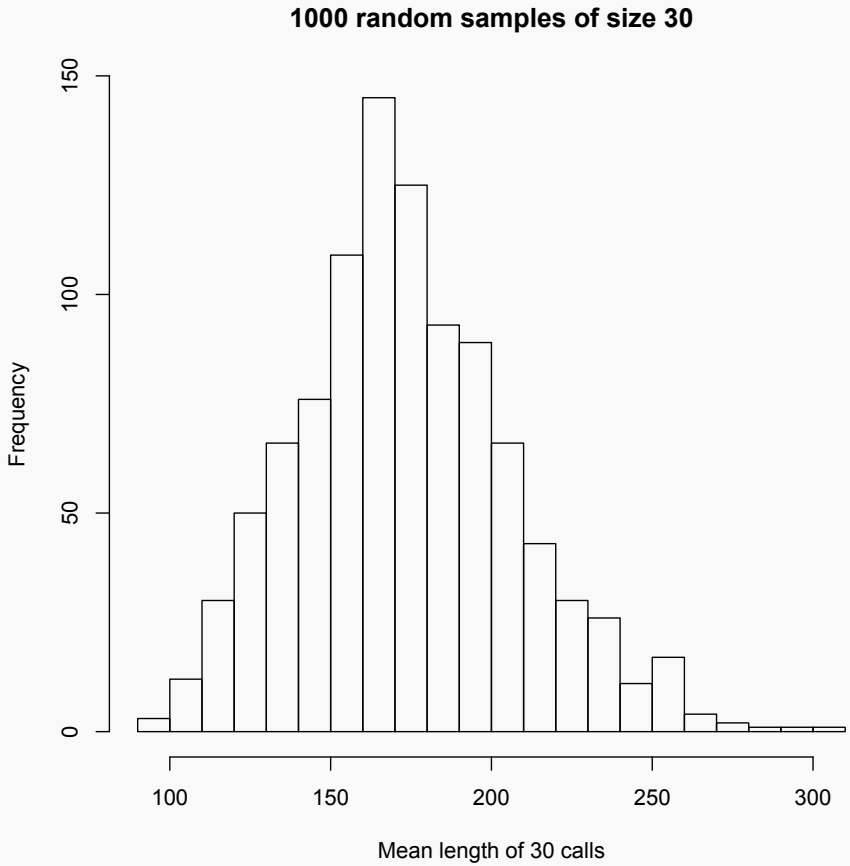


[1] Data from Moore, McCabe and Craig, *Introduction to the Practice of Statistics*, 6th ed., W.H. Freeman and Company, 2009. (outliers removed)

Central Limit Theorem (CLT)

The left plots the sample means obtained from 500 random samples of 5 calls. The right plots the sample means obtained from 500 random samples of 30 calls.



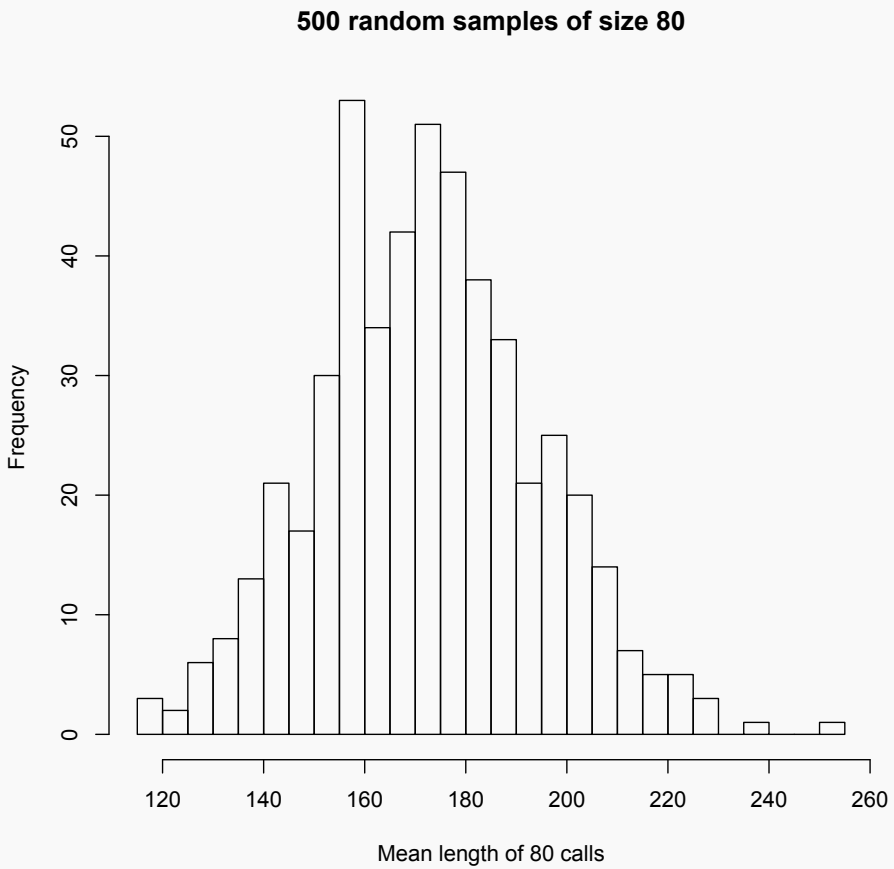


$\overline{x} = 178.4532$

$\overline{x} = 172.1415$

Central Limit Theorem (CLT)

Look what happens when we plot the sample means obtained from 500 random samples of 80 calls. Are you beginning to see the trend?



- The distribution of is approaching normality
- i.e. the histogram is looking more bell shaped and symmetrical
- The larger the n, the smaller the variance
- The mean is roughly in the same spot as the population mean ($\mu = 173.9509$)
- $\overline{x} = 172.5358$

This result is worth repeating: if n is sufficiently large then

$$\overline{X} \text{ is approximately } N(\mu, \sigma^2/n)$$

regardless of the original population!

The most troublesome part of that phrase is n is sufficiently large.

What is classified as large. Many texts books say the $n > 30$ is typically considered large.

Other Key Take Homes

1. If the original distribution is normal then the sampling distribution of the average is also normal
2. The mean of the sampling distribution is the same as the original mean
3. The standard deviation of the sampling distribistribution is smaller than the original standard deviation (gets smaller as n gets bigger)

Introduction to Statistical Inference

- Now that we know the sampling distribution of $\overline{X}$, we can begin to discuss *Statistical Inference*
- Statistical inference is the process of drawing conclusions from sample data (which are subject to random variation)
- The two types of statistical inference that we will focus on is
  - Confidence Intervals
  - Hypothesis Testing

Introduction to Confidence Intervals

Confidence intervals (CIs), provide a measure of the uncertainty that is associated with an estimate of a parameter.

For instance, the sample mean is a good estimate for the population mean (that actually gets better and better as our sample size increases), but how close can we expect $\overline{x}$ to be to μ?

Instead of quoting a single number, confidence intervals quote a range of values in interval form (a, b) to give us some idea of the variability of the estimate.

- If we had to guess one number then we would take $\overline{x}$ as our (point) estimate of μ.
- We prefer to give an interval of values in which μ is *likely* to lie (this incorporates our uncertainty)

Confidence Intervals

To construct a 95% confidence interval for μ, let me refresh you memory of a few key concepts ...

1. $\overline{X} \sim N(\mu, \sigma_{\overline{x}})$ with $\sigma_{\overline{x}} = \sigma/\sqrt{n}$ when the samples are drawn from a normal population with mean μ and variance σ

2. We can standardize this random variable such that

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{x}}} = \frac{\overline{X} - \mu}{\sigma/n} \qquad \text{where } Z \sim N(0,1)$$

3. From our tables, we know that $P(-1.96 < Z < 1.96) = 0.95$

Confidence Intervals

Using 1., 2. and 3. we can write

Confidence Intervals

Of course we can define confidence intervals for any level C.

Definition 5.4

A $100(C)\%$ confidence interval for μ is given in the form

$$\overline{x} \pm m \qquad \text{or}$$
$$(\overline{x}-m, \overline{x}+m)$$

- where m is called the **margin of error** given by $z^*\sigma/\sqrt{n}$
- z^* is the the value such that $P(-z^* < Z < z^*) = C$
- Most textbook will use $z_{\alpha/2}$ to denote the z^* defined above, where $\alpha = (1 - C)$

- When the distribution of the population is normal, this interval is exact.
- This interval is approximate for large n when the distribution of the population is not normal

Confidence Intervals

Example 5.2

Suppose we take a sample of 25 men on campus and find their average height is 1.8m. Given that the standard deviation of the population is 0.09m, give a 99% confidence interval for μ.

Confidence Intervals

- The interpretation of confidence intervals are a little tricky.
- We say that we are 99% *confident* that μ, the true unknown average height of men on campus, lies between 1.75365 and 1.84635.
- Basically, if we construct 100 different CIs based one 100 random samples, 99 of them will contain the true population mean.
- Using this interpretation, the *probability* of (1.75365, 1.84635) containing μ is either 100% if we happened to generate one of the lucky 95 and 0% if we sampled one of the unlucky 5.

Confidence Intervals

Example 5.3

Consider n =16 random draws $Y_i \sim f(y)$ with $E(Y_i) = 8$ and $Var(Y_i) = 4$. What range of values of $\overline{Y}$ would you expect to see 95% of the time?

Exercise 5.1 Consider n = 36 random draws Yi ? f (y) with E(Yi) = 4 and Var(Yi) = 16. What range of values of Y? would you expect to see 68% of the time?

Exercise 5.2 Consider a random sample of $n = 16$ points $Y_i \sim f(y)$ with $E(Y_i) = 4$ and $Var(Y_i) = 16$ what is $P(3 \leq \overline{Y} \leq 5)$

Margin of Error

As a consequence of this behaviour, as our sample size gets larger, our confidence intervals will get smaller/tighter about the mean.

This is clear once we look at the *margin of error*, $m = \frac{z^* \sigma}{\sqrt{n}}$

In words, m is a statistic expressing the amount of random sampling error in a survey's results.

Therefore, the smaller we can get this number, the better!

While decreasing the value of z^* will also achieve a decrease in m, doing so will decrease our *confidence level*.

Terminology

- A confidence interval for the mean is a range of numbers that is likely to contain μ.
- The confidence level is the percentage of confidence intervals that indeed contained our population mean had we drawn many random samples and constructed a CI for each.
- Hence, a confidence interval with confidence level .90 means that we are 90% confident that μ falls somewhere in the stated range.

Putting it into Practice

Example 5.4

We know that X comes from a normally distributed population with standard deviation 4.5. We took a sample of size 25 and calculated a sample mean of 27.3.

a) Find a 90% confidence interval for the population mean.

b) Interpret.

c) Find a 99% confidence interval for the population mean.

d) Interpret.

e) Compare the two intervals we just calculated.

a) Find a 90% confidence interval for the population mean.

b) Interpret.

c) Find a 99% confidence interval for the population mean.

d) Interpret.

e) Compare the two intervals we just calculated.

Example 5.5

What would a 100% confidence interval look like?

Significance Level

- The significance level, denoted α, is equal to 1 - (the confidence level).
- So if our confidence level is 0.94, then our significance level is 0.06.
- As our confidence level increases, our significance level decreases.

Sample Size

- Suppose we know the confidence level and margin of error that we want for our estimate (and still know the standard deviation of X).

- We can rearrange our margin of error formula to solve for the sample size we would need to achieve this

$$n = \left(\frac{\sigma * z_{\alpha/2}}{m} \right)^2$$

.

Exercises

Example 5.6

The municipal government in British Columbia is interested in the salaries of residents. They know that the standard deviation of salaries in the area is \$25,000.

a) What sample size do they need to ensure they find a 95% confidence interval that gives a margin of error of \$1,000?

b) What sample size do they need to ensure they find a 95% confidence interval that gives a margin of error of \$15,000?

c) Are there any potential issues with the last question?

a) What sample size do they need to ensure they find a 95% confidence interval that gives a margin of error of \$1,000?

b) What sample size do they need to ensure they find a 95% confidence interval that gives a margin of error of \$15,000?

c) Are there any potential issues with the last question?

Recap

- So far, we have looked confidence intervals for the population mean.
- Similar confidence intervals could be defined for other parameters such as counts and proportions (for now we will just concentrate on CI for means)
- A underlying assumption here is that we know the population, or long-run, standard deviation σ.
- This is very unrealistic. It often happens that the σ is unknown.

Introduction to Hypothesis Testing

- Hypothesis Testing is our second type of statistical inference
- This procedure is used to test claims made about the population
- We test these claims based on sample data which lead us to reject, or fail to reject the claim.
- Before we go through the details of this method, let's look at a motivating example . . .

Introduction to Hypothesis Testing

- Heinz claims that their ketchup bottle contain, on average, 20 oz. Of course, there is some inherent error at the manufacturing plant — so not all bottles will weigh exactly 20 oz.
- Say you weigh a bottle and the scale reads 19.6 oz. This might make you want to investigate whether or not their claim is true.
- So you buy 30 more bottles. If you found that the average weight in that sample was 19.88 oz, you might still suspect that Heinz is providing false advertising.
- What if the average was 19.99 oz, 19.98 oz, or 19.97 oz . . . at what average would we become concerned that Heinz was lying?

Introduction to Hypothesis Testing

- A probabilistic way of looking at it:
- Assuming that Heinz is not lying, what is the probability that I would find an average weight of only 19.88 oz in a sample of 30 bottles?
- The lower that probability, the more evidence that they might be lying.
- The higher that probability, the less evidence that they might be lying.
- This is the essence of hypothesis testing — our subject for the next while.

Hypothesis Testing for the mean

We can construct a hypothesis test to decide whether or not a hypothesis (i.e. a claim about the population parameters) is supported by our observed data (i.e. a sample from that population)
For hypothesis test about the population mean, we will check whether or not the sample mean $\overline{x}$ is (statistically) significantly different from the claimed population mean μ.

Hypothesis Testing for the mean

- If we end up with an $\bar{x}$ that is extremely unlikely under the assumed μ, then it is safe(-ish) to reject H_0.
- If we end up with an $\bar{x}$ that is likely under the assumed μ, then we should NOT reject H_0.

Hypothesis Testing for the mean

Definition 5.5

A **null hypothesis** (denoted H_0) is the statement or claim made about the population parameters being tested.

The alternative hypothesis usually states what we hope/suspect/claim to be true.

Definition 5.6

A **alternative hypothesis** is a rival statement usually denoted H_a or H_1 (we will use H_1).

Hypothesis Testing for the mean

Since we will be testing claims on the population, the null and alt. hypotheses are written in terms of population parameters. (e.g. μ) and not sample statistics (e.g. $\overline{x}$)

By carrying out a hypothesis test, we are trying to decide whether there is enough evidence to reject the null hypothesis.

As we will discuss in more detail later, we can never prove that our null hypothesis is true.

Hypothesis Testing for the mean

Although our hypotheses can be written in words, we typically express it in symbols. The hypotheses could be stated as

$$H_0 : \mu = \mu_0 \qquad \begin{cases} H_1 : \mu \neq \mu_0 & \text{two-sided test} \\ H_1 : \mu < \mu_0 & \text{one-sided (lower-tail) test} \\ H_1 : \mu > \mu_0 & \text{one-sided (upper-tail) test} \end{cases}$$

where μ_0 is some constant value.

Hypothesis Testing for the mean

The direction of our alternative hypothesis will depend on the situation at hand.
Consider our Heinz ketchup example:

- consumers of Heinz ketchup may be considered with the lower-tailed test (i.e. are we getting ripped off and getting less than we paid for?)
- Heinz corporation, on the other hand, would be more concerned with the upper-tailed test (i.e. is the company supplying more than they need to?)

Use this alternative if you no suspicion of specific direction (i.e. $>$ or $<$ than) in advanced.

Hypothesis Testing for the mean

Example 5.7

State the one-sided hypotheses for testing whether the mean weight Heinz ketchup bottles is less than 20 oz.

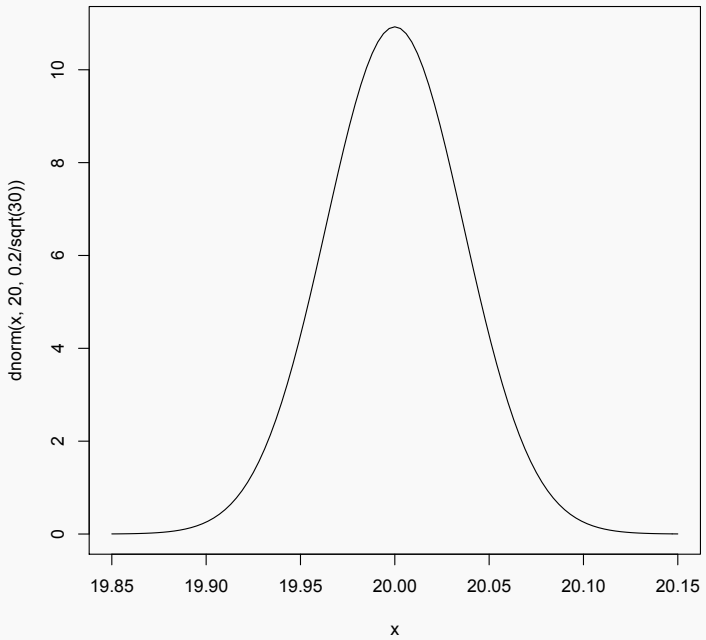
H_0: H_1:

in words . . .

In order to test the claim that the average weight is 20 oz, we suppose that H_0 is true, and see whether our sampled data provide evidence to suggest otherwise.
In other words, we find the probability that they are lying assuming that they are telling the truth.
These test will be constructed using our knowledge of the sampling distribution of $\overline{X}$.

Example 5.8

The average weight of Heinz ketchup bottles among a sample of 30 bottles was 19.88 oz. How likely is it that we observe a sample mean of 19.88 (or lower) if the true mean is actually 20 oz? (assume $\sigma = 0.2$ oz).



- In other words, we could expect a sample mean of 19.88 or lower, 5 times out of 10,000!
- This means we either
  1. We have observed a highly unlikely sample
  2. The mean weight of ketchup bottles is not actually 20 oz.

Hypothesis Testing Framework

1. State the null and alternative hypotheses
2. Calculate your test statistic
3. Make a decision based on critical values or p-values
4. State the decision in the context of the original problem

Hypothesis Testing for the mean

For hypothesis test concerning the population mean where σ is known, our **test statistic** is given by

$$Z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}},$$

where μ_0 is our hypothesized value of μ.
From our work with sampling distributions, we know that this follows a standard normal distribution $N(0, 1)$.
Once we have taken our sample, we will often referred to the realized value of Z as z_{obs}.

Hypothesis Testing for the mean

Definition 5.7 (p-value)

The p-**value** is defined as the probability that our test statistic would take on a value as extreme or more extreme than the one we actually observed, given the null hypothesis is true.

In other words, say we take our sample and find that our test statistic is equal to z_{obs}. Our p-value would be

$$\begin{cases} 2P(Z \geq |z_{obs}|) & \text{if } H_1 : \mu \neq \mu_0 \\ P(Z \leq z_{obs}) & \text{if } H_1 : \mu < \mu_0 \\ P(Z \geq z_{obs}) & \text{if } H_1 : \mu > \mu_0 \end{cases}$$

Hypothesis Test for the mean

Intuitively, the smaller the p-value, the stronger the evidence against H_0.

In order make any decisions based on our data (eg. should I stop buying Heinz ketchup), I will need to first decide small this p-value needs to get before I stop believing the claim.
We define a cut-off point α to make our final decision about H_0.

- If the p-value is $\leq \alpha$ we **reject** the null hypothesis.
- Otherwise we **fail to reject** the null hypothesis.

Hypothesis Test for the mean

We call α our significance level (or "alpha"-level) which indicates how unusual or unlikely a test statistic must be, given the null hypothesis, before we would reject the null hypothesis.
A significance level of 5% is often used but others are also common. (Unless stated otherwise, always assume α=0.05)

We say that the data is **statistically significant at level** α if the *p*-value $\leq \alpha$.

Example 5.9

In reference to Example 5.8, what conclusion can we make at a 5% significance level?

Here are a few ways we could state our conclusion:

- Since p-value is less than our significance level ($0.0005 < 0.05$), we reject our null hypothesis.
- The observed data provide very strong evidence against Heinz's claim that the mean weight of ketchup bottles is 20 oz.
- The result is statistically significant at the $\alpha = 0.05$ level.

Summary for Hypothesis Testing for the mean

1. *State the null and alternative hypothesis*

$$H_0 : \mu = \mu_0 \qquad \begin{cases} H_1 : \mu \neq \mu_0 & \text{two-sided test} \\ H_1 : \mu < \mu_0 & \text{one-sided (lower-tail) test} \\ H_1 : \mu > \mu_0 & \text{one-sided (upper-tail) test} \end{cases}$$

2. *Compute the test statistic* $z_{obs} = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$.

3. *Calculated the p-value*

$$\begin{cases} 2P(Z \geq |z_{obs}|) & \text{if } H_1 : \mu \neq \mu_0 \\ P(Z \leq z_{obs}) & \text{if } H_1 : \mu < \mu_0 \\ P(Z \geq z_{obs}) & \text{if } H_1 : \mu > \mu_0 \end{cases}$$

4. *State the conclusion*

$$\begin{cases} \text{If the } p\text{-value} \leq \alpha & \text{we } \textbf{reject} \text{ the null hypothesis} \\ \text{Otherwise} & \text{we } \textbf{fail to reject} \text{ the null hypothesis} \end{cases}$$

Rope Example I

Example 5.10

It has been established over a long period of time that the mean breaking strength of a type of rope is normally distributed with mean 41kg and standard deviation 2.4kg. A sample of 25 pieces is taken and has a mean breaking strength of 39kg; is this figure consistent with the population mean of 41kg?

Hypotheses:

Test Statistic:

p-value:

Conclusion:

Rope Example II

Example 5.10

Construct a 95% confidence interval for μ based on the information given in Example 5.10.

Two-sided Hypothesis Test and CIs

There is a one-to-one relationship between 95% CI and **two-sided** hypothesis test at a 5% significance level:

An 100*C% CI and two-sided hypothesis test at a 100*(1-C)% significance level will yield the **same conclusion!**
For example, if the hypothesized value of μ_0 falls within a 95% confidence interval, than we will fail to reject the corresponding two-sided hypothesis test with a significance level of $\alpha = 0.05$.
Alternatively, if the hypothesized value of μ_0 does not fall within a 95% confidence interval, than we will reject the corresponding two-sided hypothesis test with a significance level of $\alpha = 0.05$.

Rope Example III

A 95% Confidence Interval for μ in the Rope Example is given by:

$$39 \pm 1.96\frac{2.4}{\sqrt{25}} = 39 \pm 0.94 = (38.06, 39.94).$$

So based on the observed data, we are 95% confident that the population mean lies between 38.06kg and 39.94kg.

The fact that the hypothesized value μ_0 does not lie inside the 95% confidence interval for μ, i.e. 41 does not lie in the interval (38.06, 39.94), is equivalent to rejecting the null hypothesis that $\mu = 41$ at significance level $\alpha = 0.05$.

Example 5.11

Suppose that we take a sample of 25 men on campus and the average height of men in this sample is 1.8m. Given that the standard deviation of the population is 0.09m, is this an indication that the mean height of men on campus differs from 1.77 meters? Carry out the test a significance level of 1%.

- In the previous example we calculated the 99% CI for the mean to be
  (1.754, 1.846)
- Since 1.77 does in fact lie within the 99% CI, we should fail to reject the claim that the mean height of men on campus is equal to 1.77 meters at a 1% significance level.
- Let's verify this ...

Hypotheses:

Test Statistic:

p-value:

Conclusion: