Confidence Intervals

Ryan Miller

Outline

1. The sample average as a random variable
2. Sampling distributions
3. Confidence intervals

Introduction

- Lately we've been discussing **random variables**, which represent the unknown numeric outcome of a *random process*

Introduction

- Lately we've been discussing **random variables**, which represent the unknown numeric outcome of a *random process*
- Nearly all data is the result of a *random process*
  - We don't know which cases from the population will be sampled
  - We don't know which study participants will be randomized to the treatment/control group

Introduction

- Lately we've been discussing **random variables**, which represent the unknown numeric outcome of a *random process*
- Nearly all data is the result of a *random process*
  - We don't know which cases from the population will be sampled
  - We don't know which study participants will be randomized to the treatment/control group
- Further, *any descriptive summary* of the sample data (ie: means, proportions, correlations, etc.) is the *observed outcome* of a *random variable*

The Sample Average as a Random Variable

For a sample of n cases from a population, the sample average is calculated:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

Notice how $\bar{x}$ is the sum of many different randomly selected cases

Proportions are Averages

- Now, consider a *binary categorical* variable
  - Because binary variables involve only two categories, we can map their outcomes to the numeric values of 0 and 1
  - For example, consider a coin flip, we could map the outcome "Heads" to "1" and the outcome "Tails" to "0"

Proportions are Averages

- Now, consider a *binary categorical* variable
  - Because binary variables involve only two categories, we can map their outcomes to the numeric values of 0 and 1
  - For example, consider a coin flip, we could map the outcome "Heads" to "1" and the outcome "Tails" to "0"
- By coding the outcomes using 1s and 0s, we can see that the *sample proportion* is also an average:

$$\hat{p} = \frac{1+0+1+1+0+\ldots+1}{n}$$

- If we mapped "Heads" to a value of "1", $\hat{p}$ would refer to the proportion of heads in our sample

The Distribution of the Sample Proportion

- According to the US Census, 27.5% of the adult population are college graduates
- Randomly sampling n adults represents a *random process*
  - The proportion of college graduates in a sample, $\hat{p}$, is a *random variable*

The Distribution of the Sample Proportion

- According to the US Census, 27.5% of the adult population are college graduates
- Randomly sampling n adults represents a *random process*
  - The proportion of college graduates in a sample, $\hat{p}$, is a *random variable*
- Let's explore some different outcomes of this random variable for two different sampling protocols: random samples of size $n = 10$, and random samples of size $n = 100$
  - To begin, describe the *sample space* of $\hat{p}$ in each scenario

Random Samples of size $n = 10$

- For a single random sample of size $n = 10$, there are exactly 11 different sample proportions that could occur
  - Thus, the sample space is: $\{0/10, 1/10, 2/10, \ldots, 10/10\}$

Random Samples of size $n = 10$

- For a single random sample of size $n = 10$, there are exactly 11 different sample proportions that could occur
  - Thus, the sample space is: $\{0/10, 1/10, 2/10, \ldots, 10/10\}$
- Rather doing probability calculations, we'll simulate sampling from the population (of size $n = 10$) to judge the likelihood of each outcome
  - StatKey allows us to quickly draw many samples

Random Samples of size $n = 10$

Due to the relatively small number of discrete outcomes, it's reasonable to use a table to convey a probability model for the sample proportion:

Sample Proportion ($n = 10$)	Probability
$0/10$	$40/1000 = 0.04$
$1/10$	$150/1000 = 0.15$
$2/10$	$250/1000 = 0.25$
$3/10$	$270/1000 = 0.27$
$4/10$	$190/1000 = 0.19$
...	...
$10/10$	$0/1000 = 0$

Random Samples of size $n = 100$

- For $n = 100$, it's more sensible to use a Normal probability model
- StatKey can help us determine this model
  - Any Normal model depends upon two parameters, the expected value/mean (ie: μ) and the standard deviation (ie: σ)

Random Samples of size $n = 100$

- For $n = 100$, it's more sensible to use a Normal probability model
- StatKey can help us determine this model
  - Any Normal model depends upon two parameters, the expected value/mean (ie: μ) and the standard deviation (ie: σ)
- When the random variable our interest is a *descriptive statistic*, the term **standard error** is used to describe model's standard deviation
  - Thus, "standard deviation" is used to describe variability among individual cases, while "standard error" describe the variability of an estimate

Practice

Use StatKey to simulate drawing random samples of $n = 200$ from this population (US adults, among whom 27.5% are college graduates)

1) Is the *standard error* larger or smaller than it was for samples of size $n = 100$? Was the change *linear*? (ie: is the SE half or double what it previously had been?)
2) Use the mean and standard error as the basis for a Normal model. What percentage of samples would you expect to have between 26% and 29% colleges graduates?
3) Can you find an *interval* that you'd anticipate will contain the sample proportions found in 90% of samples?

Is simulation necessary?

The idea of reporting an *interval estimate* is very useful in describing the uncertainty involved in random processes (ie: describing sample data). However, there are two important challenges to consider:

1) We typically do not know anything about the population, everything we know is based upon the sample data
2) It's not always reasonable to do thousands of simulations

John Kerrich

- John Kerrich, a South African mathematician, was visiting Copenhagen in 1940
- When Germany invaded Denmark he was sent to an internment camp, where he spend the next five years
- To pass time, Kerrich conducted experiments exploring sampling and probability theory
  - One of these experiments involved flipping a coin 10,000 times

Kerrich's Experiment and Probability

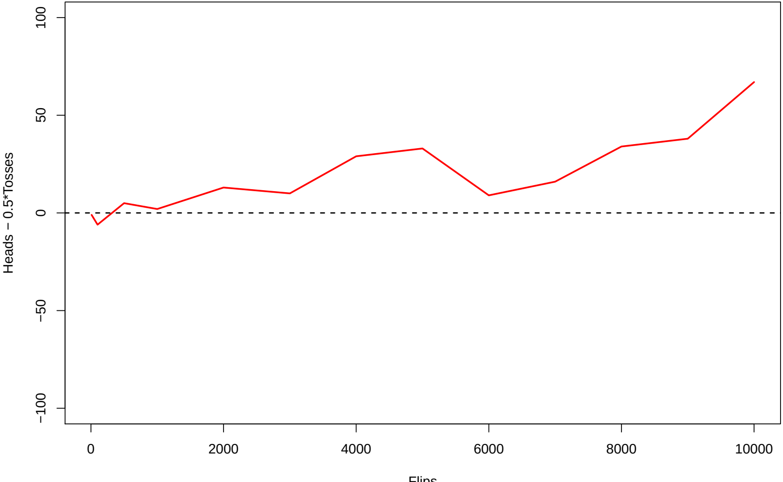
- We know that a fair coin shows "Heads" with a probability of 50%
- So, in a random sample of n coin flips, we'd expect roughly even numbers of "Heads" and "Tails"
  - We'll explore the results of Kerrich's experiment to see why the *sample average* is so special

Kerrich's Results

Number of Tosses (n)	Number of Heads	Heads - 0.5*Tosses
10	4	-1
100	44	-6
500	255	5
1,000	502	2
2,000	1,013	13
3,000	1,510	10
4,000	2,029	29
5,000	2,533	33
6,000	3,009	9
7,000	3,516	16
8,000	4,034	34
9,000	4,538	38
10,000	5,067	67

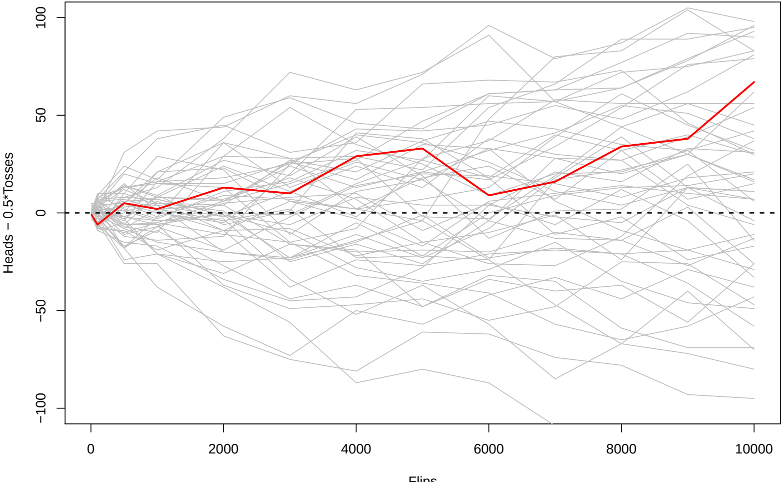
Kerrich's Results

It seems like the number of heads and tails are actually getting further apart... could this be a fluke?



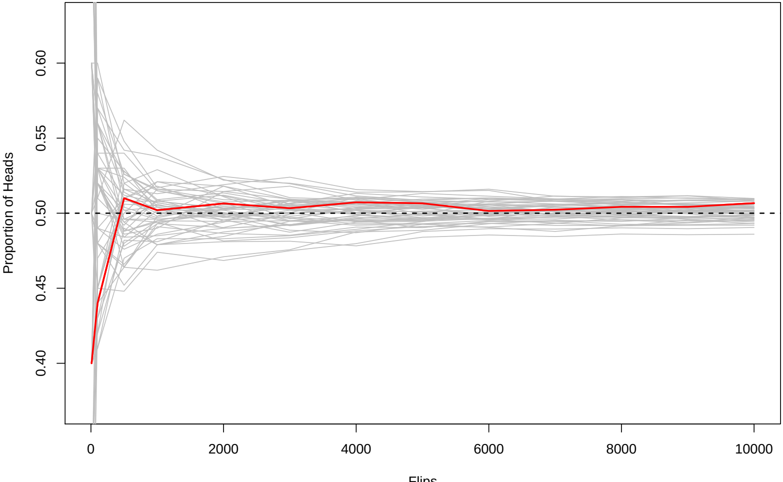
Kerrich's Experiment (repeated 50 times)

No, the phenomenon occurs systematically when repeating Kerrich's experiment



Kerrich's Experiment (sample proportions)

The *sample proportion* of heads behaves exactly as we'd expect, but why?



Central Limit Theorem

- Suppose $X_1, X_2, \dots, X_n$ are independent random variables with a common expected value $E(X)$ and variance $Var(X)$ (see previous notes for definitions of these two terms)
- Let $\bar{X}$ denote the average of all n random variables, **Central Limit Theorem** (CLT) states:

$$\sqrt{n}\left(\frac{\bar{X} - E(X)}{\sqrt{Var(X)}}\right) \to N(0, 1)$$

Central Limit Theorem

- Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with a common expected value $E(X)$ and variance $Var(X)$ (see previous notes for definitions of these two terms)
- Let $\bar{X}$ denote the average of all n random variables, **Central Limit Theorem** (CLT) states:

$$\sqrt{n}\left(\frac{\bar{X} - E(X)}{\sqrt{Var(X)}}\right) \rightarrow N(0, 1)$$

- Often it is more useful to think of CLT in the following way (which abuses notation):

$$\bar{X} \sim N\left(E(X), \frac{SD(X)}{\sqrt{n}}\right)$$

Central Limit Theorem and Sample Proportions

- The sample proportion is comprised of n different binary variables (taking on values of 1 and 0)
  - Each one of these binary variables has the same expected value and variance

Central Limit Theorem and Sample Proportions

- The sample proportion is comprised of n different binary variables (taking on values of 1 and 0)
  - Each one of these binary variables has the same expected value and variance
  - $E(X) = p * 1 + 0 * (1 - p) = p$
  - $Var(X) = p * (1 - p)^2 + (1 - p) * (0 - p)^2 = p * (1 - p)$

Central Limit Theorem and Sample Proportions

- The sample proportion is comprised of n different binary variables (taking on values of 1 and 0)
  - Each one of these binary variables has the same expected value and variance
  - $E(X) = p * 1 + 0 * (1 - p) = p$
  - $Var(X) = p * (1 - p)^2 + (1 - p) * (0 - p)^2 = p * (1 - p)$
- Thus, the *sampling distribution* of sample proportions is:

$$\hat{p} \sim N(p, \sqrt{p(1 - p)/n})$$

The Power of CLT

- Central Limit Theorem is one of the most important theoretical results in all of statistics
- In real-world applications, it is nearly impossible to know the probability distribution of something that is only observed once (remember that real researchers can only afford to collect a single sample)

The Power of CLT

- Central Limit Theorem is one of the most important theoretical results in all of statistics
- In real-world applications, it is nearly impossible to know the probability distribution of something that is only observed once (remember that real researchers can only afford to collect a single sample)
- But by focusing on the *sample average* this isn't an issue, as CLT provides us the distribution of sample averages
  - That is, we are able to use CLT to understand the *sampling variability* of our study, despite only getting to see a single sample!

Example

- Let's consider a random sample of $n = 100$ coin flips
  - What proportion of heads might we expect? It'll likely be close to 50%, but we know there's sampling variability, the question is how much...

Example

- Let's consider a random sample of $n = 100$ coin flips
  - What proportion of heads might we expect? It'll likely be close to 50%, but we know there's sampling variability, the question is how much...
- Each coin flip is a random variable an expected value of 0.5, so Central Limit Theorem tells us that proportion of heads in random samples of $n = 100$ coin flips follows a Normal distribution:

$$\hat{p} \sim N(0.5, \sqrt{0.5(1-0.5)/100})$$

Assumptions

Using the Central Limit theorem to determine the distribution of sample averages is only appropriate when the following conditions are met:

1) *Independence* - the cases in the sample (ie: the individual contributions to the sample average) are not related to each other
2) *Large population* - less that 10% of the population is being sampled (otherwise removing the already sampled individuals has too much of an impact on the probability of selection)
3) *Large sample* - $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$

Most of the time, it's only the third condition that is problematic

Practice

Let's revisit the selection of a random sample of size $n = 100$ from the population of US adults (of which 27.5% are college graduates)

1) Use Central Limit theorem to come up with a Normal model for the proportion of college graduates in the sample.
2) How does the *standard error* in this model compare to the standard error we found by simulation? (recall it was approximately 0.045)

Statistical Inference

Recall that the *fundamental goal* of statisticians is to use information from a sample to make *reliable* statements about a population, a process known as **statistical inference**:

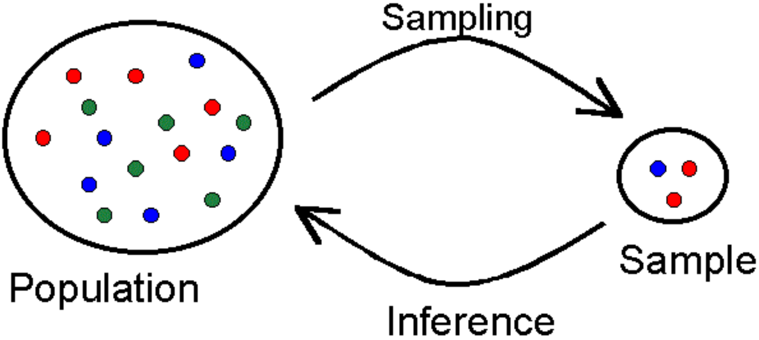


Image credit: http://testofhypothesis.blogspot.com/2014/09/the-sample.html

Point Estimation

- If a sampling protocol is *unbiased*, the sample average is a sensible estimate of the population mean
  - This is called a **point estimate**, referring to the fact that it is a single value

Point Estimation

- If a sampling protocol is *unbiased*, the sample average is a sensible estimate of the population mean
  - This is called a **point estimate**, referring to the fact that it is a single value
- From our study of *sampling distributions*, we know that the existence of sampling variability means a point estimate is almost certainly wrong (at least to some degree)
  - This suggests that we can more appropriately describe what we think is true of the population by reporting an **interval estimate** that accounts for *sampling variability*

Point vs. Interval Estimation

To summarize:

- **Point estimation** uses sample data to produce a *single "most likely" estimate* of a population characteristic, which will almost always miss the target (at least by some degree)
- **Interval estimation** uses sample data to produce a *range of plausible estimates* of a population characteristic, an approach that has a much better chance at capturing the truth

Point vs. Interval Estimation

To summarize:

- **Point estimation** uses sample data to produce a *single "most likely" estimate* of a population characteristic, which will almost always miss the target (at least by some degree)
- **Interval estimation** uses sample data to produce a *range of plausible estimates* of a population characteristic, an approach that has a much better chance at capturing the truth

An analogy:

Using only a point estimate is like fishing in a murky lake with a spear. We can throw a spear where we saw a fish, but we will probably miss. On the other hand, if we toss a net in that area, we have a good chance of catching the fish.

Margin of Error

Most interval estimates have the form:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

We often report these intervals using only their endpoints:

$$(\text{Est} - \text{MOE}, \text{Est} + \text{MOE})$$

Margin of Error

Most interval estimates have the form:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

We often report these intervals using only their endpoints:

$$(\text{Est} - \text{MOE}, \text{Est} + \text{MOE})$$

- We'd like the *margin of error* to be constructed in way that carries a *quantifiable* claim of precision
  - ie: 80% of the time an interval with this type of margin of error will contain the population characteristic
  - Without an accompanying claim regarding precision, reporting a margin of error is not particularly useful

Practice

So, what can we say about a population proportion, p, based upon an observed sample proportion, $\hat{p}$? Consider a representative sample of 100 infants used to estimate the proportion of all babies who are born prematurely

- True or false? "We observed $\hat{p} = 0.14$, so we know that 14% of all babies are born prematurely"
- True or false? "We observed $\hat{p} = 0.14$, it's probably true 14% of all babies are born prematurely"
- True or false? "Although we don't know p, if we attach a large margin error to our point estimate,the interval estimate $14\% \pm 10\% = (4\%, 24\%)$ probably contains p"

Confidence intervals

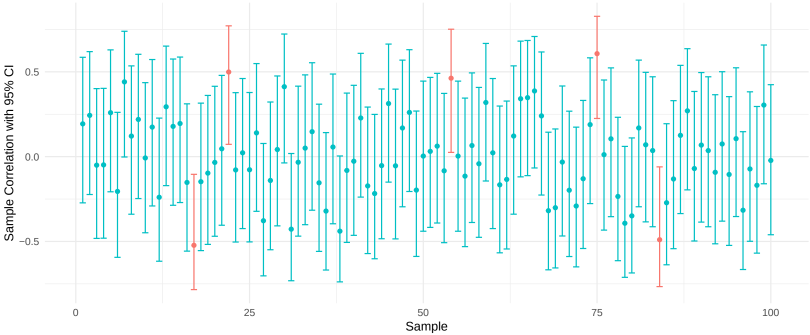
- A **confidence interval** is an interval estimate whose margin of error is based upon a procedure with a long-run "success rate" known as a *confidence level*

Confidence intervals

- A **confidence interval** is an interval estimate whose margin of error is based upon a procedure with a long-run "success rate" known as a *confidence level*
  - A 95% confidence interval was created using a procedure that will succeed in containing the true population parameter in 95% of different random samples (or study replications)
  - The confidence level does not describe the likelihood that particular interval succeeds, instead it describes the estimation procedure's long-run success rate

Confidence intervals

Shown below are 95% CI estimates from 100 different random samples ($n = 20$) drawn from a population with correlation of $\rho = 0$



Notice that 5 of 100 samples resulted in a 95% CI that failed to contain the true population-level correlation!

Confidence Intervals and the Normal model

Under a *Normal probability model*, confidence intervals can be found via:

$$\text{Point Estimate} \pm c * SE$$

- "c" is a percentile of the *Standard Normal distribution* that is used to calibrate the interval (sometimes called $z*$)
  - For example, $c = 1.96$ is used for a 95% CI
- SE is found using the results of the Central Limit theorem
  - For example, for the sample proportion $\hat{p} = 12/30 = 0.4$, $SE = \sqrt{\frac{0.4*(1-0.4)}{30}} = 0.09$

Practice

A 2021 study looked at the true-positive rate (sensitivity) of an Abbott Diagnostics rapid test for Covid-19 (one of the earliest such tests). Of the 84 cases with symptomatic Covid-19 who took the test, 38 had a "positive" result. Our goal is to estimate p, the overall sensitivity of this test in the target population (ie: all symptomatic Covid cases).

1) Verify that the conditions are met to use the CLT Normal approximation to construct a confidence interval estimate
2) Find the values of $\hat{p}$, its SE, and the value of c needed to construct a 99% CI estimate of p
3) Calculate and interpret the 99% CI

Practice (solution)

1) First, $\hat{p} = 38/84 = 0.452$. Then, $n\hat{p} = 84 * 0.452 = 38$ and $n(1-\hat{p}) = 84 * (1 - 0.452 = 46$. Because both are larger than 10, the CLT Normal approximation is reasonable.
2) $\hat{p} = 38/84 = 0.452$, $SE = \sqrt{\frac{0.452(1-0.452)}{84}} = 0.054$, and $c = 2.576$ (this defines the middle 99% of a N(0,1) curve)
3) The 99% CI is $0.452 \pm 2.576 * 0.054 = (0.313, 0.591)$. Our sample suggests, with 99% confidence, that the true sensitivity of the Abbott rapid test is somewhere between 31.3% and 59.1%

CLT Standard Errors

Without getting into the details, the CLT standard error formulas for commonly used descriptive statistics are shown below:

- Single proportion, $\hat{p}$, $SE = \sqrt{\frac{p(1-p)}{n}}$
  - works well when $n * p \geq 10$ and $n * (1 - p) \geq 10$
- Difference in proportions, $\hat{p}_1 - \hat{p}_2$, $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - works well when $n_i * p_i \geq 10$ and $n_i * (1 - p_i) \geq 10$ for both groups
- Single mean, $\bar{x}$, $SE = \frac{\sigma}{\sqrt{n}}$
  - works well when $n \geq 30$
- Difference in means, $\bar{x}_1 - \bar{x}_2$, $SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
  - works well when $n_i \geq 30$ for both groups
- Correlation coefficient, r, $SE = \sqrt{\frac{1-\rho^2}{n-2}}$
  - works well when $n \geq 30$

Note: using any of these formulas requires the replacement of unknown population parameters.

Practice #1

The previously mentioned study also examined a test produced by Siemens. Of 72 cases with symptomatic Covid-19 that took the Siemens test, 39 had a "positive" result. Recall that 38 of 84 symptomatic cases tested positive on the Abbott test. Our goal is estimate $p_1 - p_2$, the difference in sensitivity of these two tests (at the population level)

1) Let $\hat{p}_1 = 38/84 = 0.45$ be the sample proportion for the Abbott test, and $\hat{p}_2 = 39/72 = 0.54$ be the sample proportion for the Siemens test. Find the SE for the difference in proportions, $\hat{p}_1 - \hat{p}_2$.
2) Find and interpret the 95% CI estimate of $p_1 - p_2$

Practice #1 (solution)

1) $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{0.45(1-0.45)}{84} + \frac{0.54(1-0.54)}{72}} = 0.08$
2) Since $c = 1.96$, $\hat{p}_1 - \hat{p}_2 = 0.54 - 0.45 = -0.09$, and $SE = 0.08$, we calculate: $-0.09 \pm 1.96 * 0.08 =$ (-0.247, 0.067). This represent the plausible range of differences in sensitivity of these tests at the population level (estimated with 95% confidence). Because zero is included in this interval, it's plausible that the tests are actually no different.

Practice #2

A random sample of $n = 33$ adult women found a correlation coefficient of $r = 0.68$ between bodyweight and resting metabolic rate.

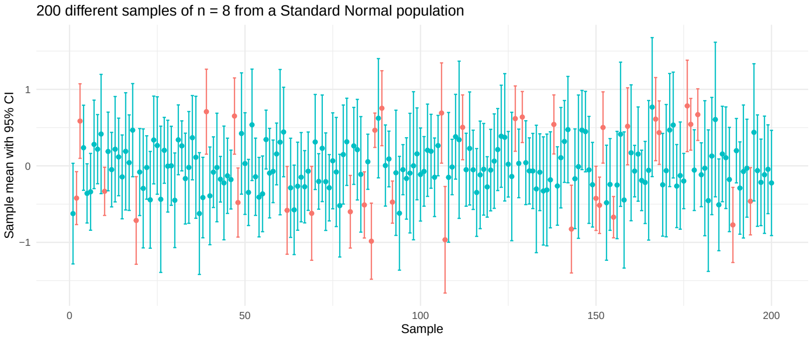
1) Can you be statistically confident that bodyweight is resting metabolic rate using only the *point estimate*?
2) Find and interpret the 99% CI estimate of ρ (the population-level correlation)

Practice #2 (solution)

1) No, even though a strong correlation was seen in the sample, the data consist of only 33 cases, so that correlation might be explained by sampling variability
2) $c = 2.576$ for 99% confidence and $SE = \sqrt{\frac{1-0.68^2}{30-2}} = 0.139$; so the 99% CI is $0.68 \pm 2.576 * 0.139$ or (0.322, 1.00) - note that we'd want to cap the upper endpoint at 1.00 (as that's the maximum a correlation coefficient can be)

William Gosset and the t-distribution

- Using the CLT to calculate confidence intervals for quantitative outcomes (means, differences in means) requires an estimate of a second unknown population parameter (σ)
  - If we use the corresponding sample estimate, s, this is what happens:



William Gosset and the t-distribution

- Clearly the prior 95% CI procedure is *invalid* - too many of the intervals didn't contain μ
- William Gosset, a chemist working for Guinness Brewing, became aware of this issue in the 1890s
  - His work evaluating the yields of different barley strains frequently involved small sample sizes

William Gosset and the t-distribution

- Clearly the prior 95% CI procedure is *invalid* - too many of the intervals didn't contain μ
- William Gosset, a chemist working for Guinness Brewing, became aware of this issue in the 1890s
  - His work evaluating the yields of different barley strains frequently involved small sample sizes
- In 1906, Gosset took a leave of absence from Guinness to study under Karl Pearson (developer of the correlation coefficient)
  - Gosset discovered the issue was due to using *s* interchangeably with σ

William Gosset and the t-distribution

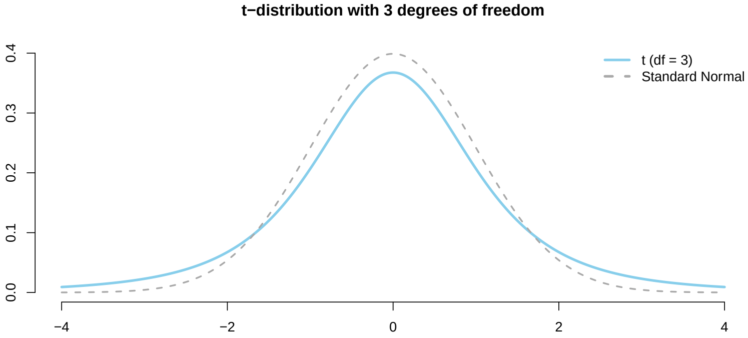
- Treating s as if it were a perfect estimate of σ results in a systematic underestimation of the total amount of variability involved in making the CI
  - To account for the additional variability introduced by estimating σ using s, a modified distribution that's slightly more spread out than the Standard Normal curve must be used

William Gosset and the t-distribution

- Treating s as if it were a perfect estimate of σ results in a systematic underestimation of the total amount of variability involved in making the CI
    - To account for the additional variability introduced by estimating σ using s, a modified distribution that's slightly more spread out than the Standard Normal curve must be used
- Typically the inventor of a new method gets to name it after themselves
    - However, Gosset was forced to publish his new distribution under the pseudonym "student" because Guinness didn't want it's competitors knowing they employed statisticians!
    - Student's t-distribution is now among the most widely used statistical results of all time

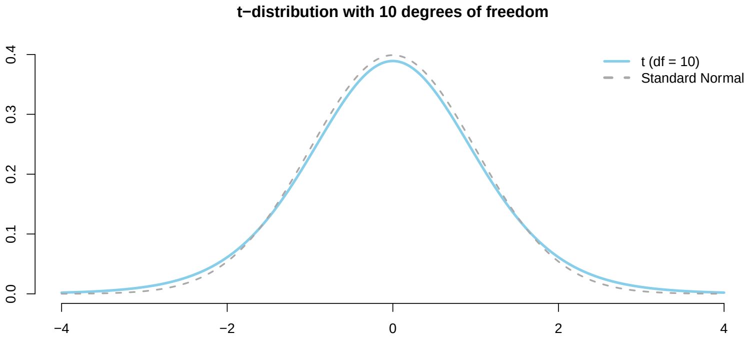
The t-distribution

The *t*-distribution accounts the additional uncertainty in small samples using a parameter known as *degrees of freedom*, or *df*:

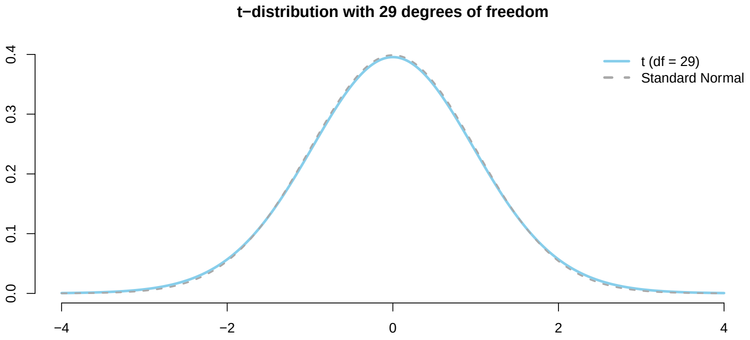


When estimating a single mean, $df = n - 1$

The t-distribution



The t-distribution



Practice

While waiting at an airport, a traveler notices 6 flights to similar a similar part of the country were delayed 6, 10, 13, 23, 45, 55 minutes. The mean delay in this sample was 25.33, with a sample standard deviation of $s = 20.2$. Assuming these data are a representative sample, answer the following:

1) How many degrees of freedom are involved when using the t-distribution to form a CI estimate? What is the value of c that should be used for 95% confidence?
2) What is the 95% CI estimate for the average delay of flights to the part of the country this traveler is heading?

Practice (solution)

1) Because $n = 6$, we'd use $df = n - 1 = 5$. For $df = 5$, $c = 2.571$ defines the middle 95% of the distribution.
2) Point Estimate $\pm$ MOE, Point estimate $= \bar{x} = 25.33$, Margin of error $= c * SE = 2.571 * \frac{20.2}{\sqrt{6}}$
   - All together, 95% CI: $25.33 \pm 2.571 * \frac{20.2}{\sqrt{6}} = (4.1, 46.5)$
   - We are 95% confident the *average* delay is somewhere between 4.1 minutes and 46.5 minutes

Note: if we'd erroneously used a Normal model (instead of the *t*-distribution), we'd get an interval that is much narrower (9.2, 41.5), but this interval wouldn't have the confidence level we are advertising (ie: it wouldn't really be a 95% CI because it would miss too often)

When to use the t-distribution

- The t-distribution was designed for small, Normally distributed samples
  - However, it can also be reliably used on large samples, regardless of their shape

	Sample data are approximately Normal	Sample data are non-Normal or skewed
Sample size is large ($n \geq 30$)	Use t-distribution	Use t-distribution
Sample size is small ($n < 30$)	Use t-distribution	*do not* use t-distribution

- For small, non-Normal samples, more robust methods (such as bootstrapping) should be used instead

Conclusion

- Confidences intervals are preferred over point estimates because they address the question of statistical uncertainty that is inherent to random processes (such as data collection, random assignment, etc.)
- A confidence interval provides a plausible range of values for an unknown population parameter
  - The confidence level describes the success rate of the method used to calculate the interval if it were applied to many random samples
- To find a confidence interval you need three components:
  - A point estimate (calculated from the sample data)
  - The *SE* of that estimate (found using a CLT formula)
  - A calibration constant, c (found using a Normal distribution or a t-distribution)
- The t-distribution should be used for a single mean or a difference in means