Math 11
Calculus-Based Introductory Probability and Statistics

Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today:

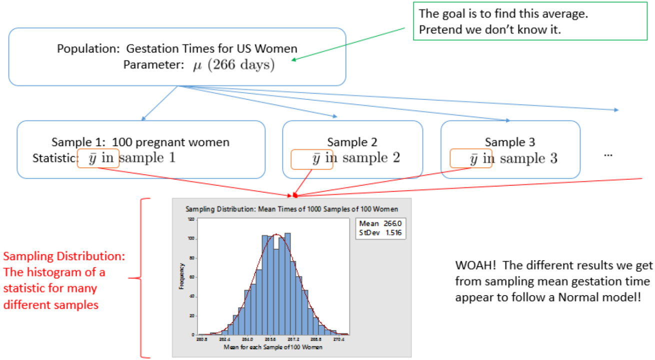
- Hypothesis Testing for Means (t-test)
- Better Understanding Errors Made in Hypothesis Tests

Inference for Means

We want to do all the same inference for means that we did with proportions: understanding the sampling distribution, hypothesis testing, and confidence intervals

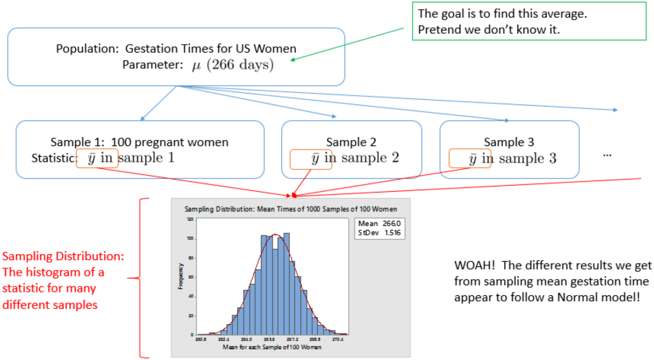
Inference for Means

We want to do all the same inference for means that we did with proportions: understanding the sampling distribution, hypothesis testing, and confidence intervals



Inference for Means

We want to do all the same inference for means that we did with proportions: understanding the sampling distribution, hypothesis testing, and confidence intervals



In a previous class, we learned that the sampling distribution is approximately Normal with

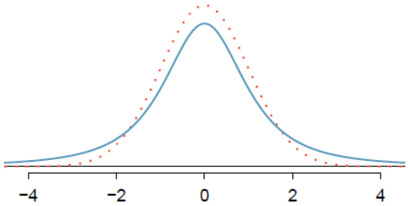
$$\mu_{model} = \mu_{population} \qquad \text{and} \qquad \sigma_{mode} = \frac{\sigma_{model}}{\sqrt{n}}.$$

Time for the Truth...

With smaller sample sizes (often $n < 30$) or populations where you don't know σ (and must approximate it using s_x), there is a better approximation of the sampling distribution than the Normal model.

Time for the Truth...

With smaller sample sizes (often $n < 30$) or populations where you don't know σ (and must approximate it using s_x), there is a better approximation of the sampling distribution than the Normal model.

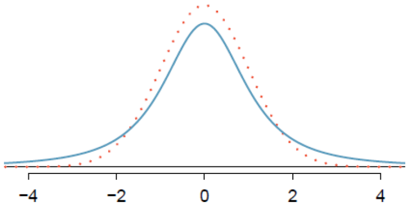


Solid line: t-distribution
Dotted line: Normal (z) distribution

Like the Normal distribution, the t-distribution is unimodal and symmetric.
Note, however, that the tails of the t-distribution are thicker, and this can change the area under the curve, and hence, the p-values that are based on it.

Time for the Truth...

With smaller sample sizes (often $n < 30$) or populations where you don't know σ (and must approximate it using s_x), there is a better approximation of the sampling distribution than the Normal model.



Solid line: t-distribution
Dotted line: Normal (z) distribution

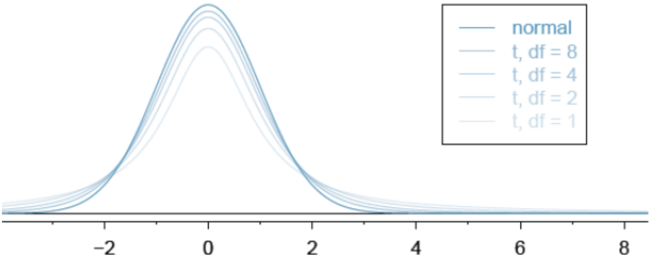
Like the Normal distribution, the t-distribution is unimodal and symmetric.
Note, however, that the tails of the t-distribution are thicker, and this can change the area under the curve, and hence, the p-values that are based on it.

When Should I Use Z vs. T?

- If you know σ (almost never true): Use a z-distribution.
- In all other cases: Use a t-distribution.

From z to t

In practice, we tend **not** to use the Normal curve as the approximation of the sampling distribution because the t-distribution gives us more precise results.



There are many curves in the t-distribution family. You choose the appropriate one based on the size of your sample.

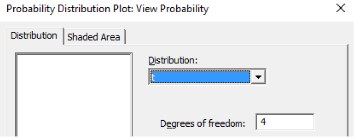
With n data points in the sample, you use the t-distribution with
$$df = n - 1.$$

df stands for "degrees of freedom" and as it gets bigger, the t-distribution morphs into a standard normal distribution.

Welcome to the New World

How much area is to the right of -0.6 on t_4?

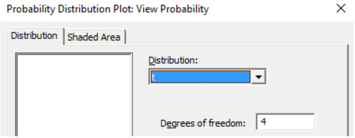
As before: Graph >> Probability Distribution Plot
>> View Probability

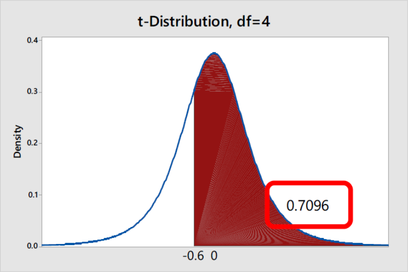


Welcome to the New World

How much area is to the right of -0.6 on t_4?

As before: Graph >> Probability Distribution Plot
>> View Probability

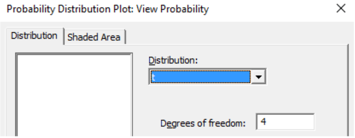


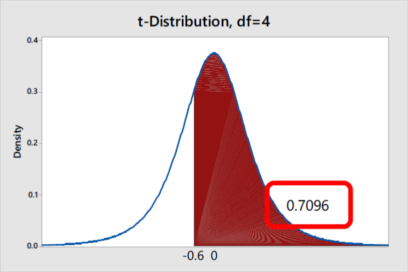


Welcome to the New World

How much area is to the right of -0.6 on t_4?

As before: Graph >> Probability Distribution Plot
>> View Probability



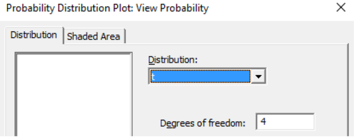


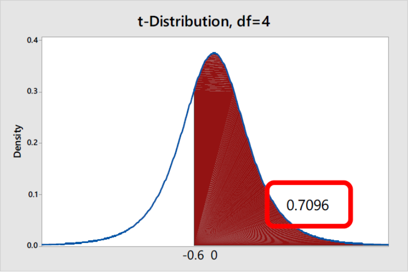
What percent of the t_6 distribution is more than 2 unites from 0?

Welcome to the New World

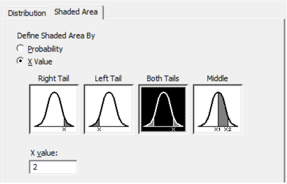
How much area is to the right of -0.6 on t_4?

As before: Graph >> Probability Distribution Plot
>> View Probability





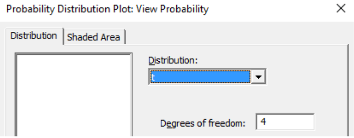
What percent of the t_6 distribution is more than 2 unites from 0?

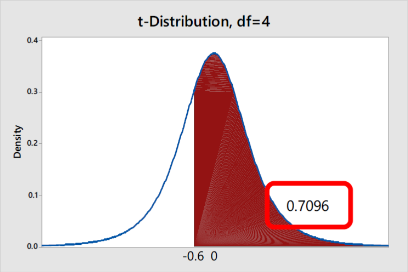


Welcome to the New World

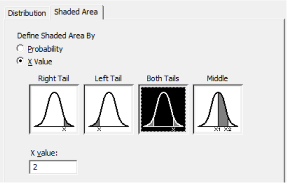
How much area is to the right of -0.6 on t_4?

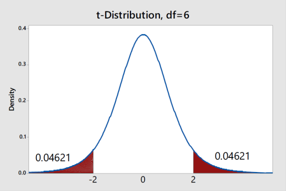
As before: Graph >> Probability Distribution Plot
>> View Probability





What percent of the t_6 distribution is more than 2 unites from 0?





$2 \times 0.046 = 0.092$.

(On $N(0,1)$, this answer would be about $1 - 0.95 = 0.05$)

Using Tables with t-Distributions

Since they don't want to print a new table for every possible *df* value, they print one table, but it is not as good as a z-table (standard Normal).
Also, a t-table is usually reversed (areas on outside, critical values in the table).

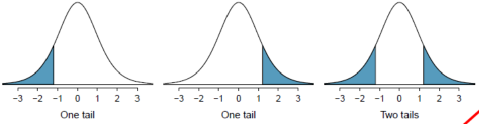


Figure B.1: Tails for the t-distribution.

See back of your book

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36

(Redoing the last problem) Our critical value of 2 is between 1.94 and 2.45.

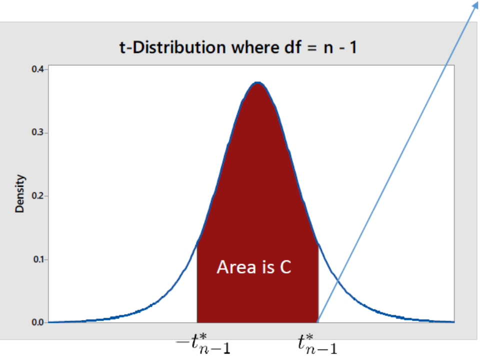
So, the area we seek is somewhere between 0.05 and 0.10. (Technology gave an answer about 0.092.)

IMPORTANT: With a T-table, you will only be able to imprecisely find the area (i.e., find bounds on the area).

New, Improved CIs for Means!

For means, our CIs will use the same setup as before, but we live on a t-distribution, not a z-distribution:

$$\bar{x} \pm t^*_{n-1} \cdot SE_{\bar{x}}$$



In a sample with n data points, the best model for the sampling distribution is t_{n-1}. We still have $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s_x}{\sqrt{n}}$

First C.I. With the t-Distribution

On average, how much do U.S. baby girls weigh? To find out, you sneak into random hospitals and collect illegally the weight of 12 random newborn babies. If $\bar{x} = 7.3$ lbs and $s_x = 2$, find a 90% C.I. for μ, the average weight of all U.S. female babies.

First C.I. With the t-Distribution

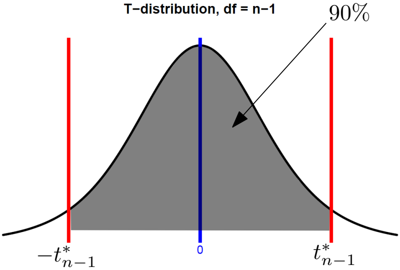
On average, how much do U.S. baby girls weigh? To find out, you sneak into random hospitals and collect illegally the weight of 12 random newborn babies. If $\bar{x} = 7.3$ lbs and $s_x = 2$, find a 90% C.I. for μ, the average weight of all U.S. female babies.

1) Decide on the model and find the critical value for your confidence level.
Since $n = 12$, our sampling distribution is modeled by $t_{n-1} = t_{11}$.

First C.I. With the t-Distribution

On average, how much do U.S. baby girls weigh? To find out, you sneak into random hospitals and collect illegally the weight of 12 random newborn babies. If $\bar{x} = 7.3$ lbs and $s_x = 2$, find a 90% C.I. for μ, the average weight of all U.S. female babies.

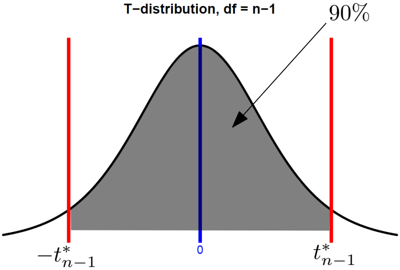
1) Decide on the model and find the critical value for your confidence level.
Since $n = 12$, our sampling distribution is modeled by $t_{n-1} = t_{11}$.



First C.I. With the t-Distribution

On average, how much do U.S. baby girls weigh? To find out, you sneak into random hospitals and collect illegally the weight of 12 random newborn babies. If $\bar{x} = 7.3$ lbs and $s_x = 2$, find a 90% C.I. for μ, the average weight of all U.S. female babies.

1) Decide on the model and find the critical value for your confidence level.
Since $n = 12$, our sampling distribution is modeled by $t_{n-1} = t_{11}$.



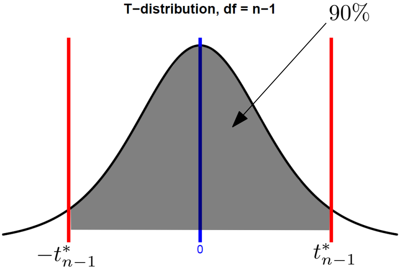
one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05

First C.I. With the t-Distribution

On average, how much do U.S. baby girls weigh? To find out, you sneak into random hospitals and collect illegally the weight of 12 random newborn babies. If $\bar{x} = 7.3$ lbs and $s_x = 2$, find a 90% C.I. for μ, the average weight of all U.S. female babies.

1) Decide on the model and find the critical value for your confidence level.

Since $n = 12$, our sampling distribution is modeled by $t_{n-1} = t_{11}$.



one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05

We read $t^* = 1.80$.

2) Break out the C.I. formula.

$$\begin{aligned}\bar{x} \pm t^*_{n-1} \times SE_{\bar{x}} &\simeq 7.3 \pm 1.8 \times \frac{2}{\sqrt{12}} \\ &= (6.26, 8.34).\end{aligned}$$

2) Break out the C.I. formula.

$$\begin{aligned}\bar{x} \pm t^*_{n-1} \times SE_{\bar{x}} &\simeq 7.3 \pm 1.8 \times \frac{2}{\sqrt{12}} \\ &= (6.26, 8.34).\end{aligned}$$

3) Interpret in plain English.

We are 90% confident that μ is between 6.26 and 8.34 pounds.

2) Break out the C.I. formula.

$$\bar{x} \pm t^*_{n-1} \times SE_{\bar{x}} \simeq 7.3 \pm 1.8 \times \frac{2}{\sqrt{12}}$$
$$= (6.26, 8.34).$$

3) Interpret in plain English.

We are 90% confident that μ is between 6.26 and 8.34 pounds.

Remark: The actual value for μ is 7.5 pounds. Our C.I. was one of the lucky 90%.

Your Turn!

You decide to build an 80% C.I. for some mean you are estimating. What is the critical value if your sample has size 15?

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81

- 1.34
- 1.35
- 1.75
- 1.76

Your Turn!

You decide to build an 80% C.I. for some mean you are estimating. What is the critical value if your sample has size 15?

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81

- 1.34
- 1.35
- 1.75
- 1.76

Answer: $t^*_{n-1} = t^*_{14} = 1.35$.

Your Turn!

About how much area is to the left of -2 on t_{15}?

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17
	11	1.36	1.80	2.20	2.72	3.11
	12	1.36	1.78	2.18	2.68	3.05
	13	1.35	1.77	2.16	2.65	3.01
	14	1.35	1.76	2.14	2.62	2.98
	15	1.34	1.75	2.13	2.60	2.95
	16	1.34	1.75	2.12	2.58	2.92
	17	1.33	1.74	2.11	2.57	2.90
	18	1.33	1.73	2.10	2.55	2.88
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81

- Between 0.1 and 0.2
- Between 0.05 and 0.1
- Between 0.025 and 0.05
- Between 0.001 and 0.025

Your Turn!

About how much area is to the left of -2 on t_{15}?

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81

- Between 0.1 and 0.2
- Between 0.05 and 0.1
- Between 0.025 and 0.05
- Between 0.001 and 0.025

Answer: Between 0.025 and 0.05

Your Turn!

Suppose we want to do inference on a mean using a t-distribution. When does a t-distribution actually model the sample distribution?

1. It always does. It doesn't matter what the sample size is or what the population distribution looks like.
2. We only need the data points in our sample to be independent
3. If the data in the sample were chosen at random and are $<10\%$ of the population
4. We need randomization, <10 % population, and the population to be reasonably normal looking (skew is ok with a larger n).

Your Turn!

Suppose we want to do inference on a mean using a t-distribution. When does a t-distribution actually model the sample distribution?

1. It always does. It doesn't matter what the sample size is or what the population distribution looks like.
2. We only need the data points in our sample to be independent
3. If the data in the sample were chosen at random and are <10% of the population
4. We need randomization, <10 % population, and the population to be reasonably normal looking (skew is ok with a larger n).

Answer: 4.

Hypothesis Testing with t-Distributions

15 years ago, the average finishing time of a Marathon was 3.683 hours. You are curious if runners are getting faster or slower in the current year.

Hypothesis Testing with t-Distributions

15 years ago, the average finishing time of a Marathon was 3.683 hours. You are curious if runners are getting faster or slower in the current year.

Let μ be the average time of finishers in the current year.

$$H_0 : \mu = 3.683 \; ; \; H_A : \mu \neq 3.683$$

Hypothesis Testing with t-Distributions

15 years ago, the average finishing time of a Marathon was 3.683 hours. You are curious if runners are getting faster or slower in the current year.

Let μ be the average time of finishers in the current year.

$$H_0 : \mu = 3.683 \; ; \; H_A : \mu \neq 3.683$$

You collect data for 20 random runners in the current year, and get $\bar{x} = 3.8$ with $s_x = 0.5$. Run a hypothesis test with $\alpha = 0.05$.

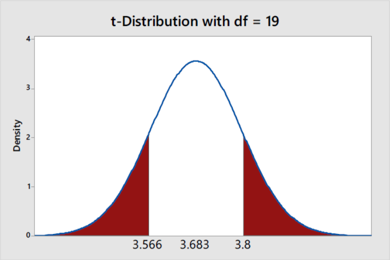
You should **always** convert to the standard t-distribution. To do so, you'll need the t-score:

$$\begin{aligned} T &= \frac{\text{point estimate} - \text{ null value}}{SE} \\ &= \frac{3.8 - 3.683}{0.5/\sqrt{20}} \simeq 1.046. \end{aligned}$$

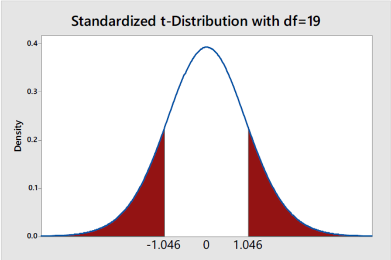
You should **always** convert to the standard t-distribution. To do so, you'll need the t-score:

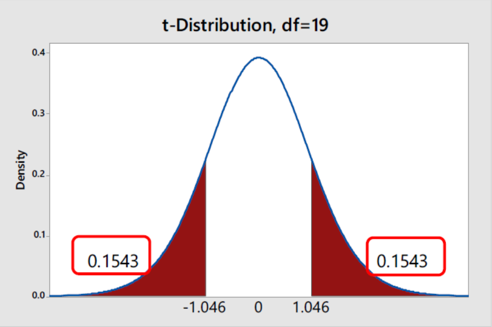
$$T = \frac{\text{point estimate} - \text{ null value}}{SE}$$
$$= \frac{3.8 - 3.683}{0.5/\sqrt{20}} \simeq 1.046.$$

Original Setting



Standardized Setting





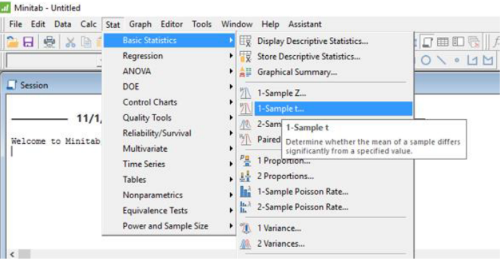
Technology says the area is 0.154 + 0.154 = 0.308

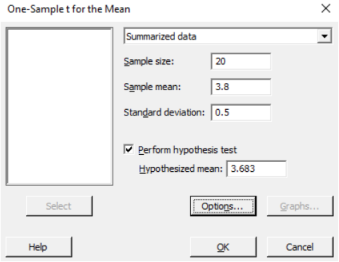
Since $P = 0.31 > 0.05$, we do not reject H_0. Our data are not strong enough to show a difference (if there is one).

In a table, our T-score of 1.046 is to the left of 1.33. So we have more area than the two-tails value of 0.2. So, P-value $> 0.2 > 0.05$, as before.

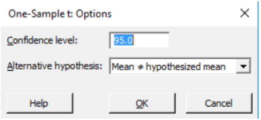
one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

The Last Example Done Easily





Under options, you need to tell Minitab what H_A is so it knows what area it is finding.



Set a one-sided H_A on this screen, if that is the case

One-Sample T

Test of μ = 3.683 vs ≠ 3.683

N	Mean	StDev	SE Mean	95% CI	T	P
20	3.800	0.500	0.112	(3.566, 4.034)	1.05	0.308

What Assumptions Must We Meet to Use t-Distributions?

What Assumptions Must We Meet to Use t-Distributions?

When doing statistical inference (using CIs or hypothesis testing), you rely on the shape and SE of the sampling distribution. To get these to be correct, you must have:

Independence of data: Knowing one piece of data should not help you predict other data. Since this is hard to check, this is usually replaced with the:

- Randomization Condition
- $< 10\%$ Condition

Population distribution must be nearly normal. Since this is hard to check:

- Look for near normality in the histogram of your sample (or its qq-plot).
- More skew is OK as n gets larger.

Beetle Study

We study beetle biodiversity in a pasture. For this, we collect a biodiversity index (Steinhaus index) in 12 parcels and get the following data:

$$\bar{x} = 0.2505 \text{ and } s_x = 0.0959.$$

Beetle Study

We study beetle biodiversity in a pasture. For this, we collect a biodiversity index (Steinhaus index) in 12 parcels and get the following data:

$$\bar{x} = 0.2505 \text{ and } s_x = 0.0959.$$

An environmental engineer, specialist of beetle populations, tells you than an average biodiversity index lower than 0.49 in the pasture would be worrying.

Build a test with level of confidence $\alpha = 5\%$ for determining if the state of the pasture is worrying.

Beetle Study

1) We build hypotheses for the situation

Beetle Study

1) We build hypotheses for the situation

$$H_0\colon\ \mu = 0.49,$$
$$H_A\colon\ \mu < 0.49.$$

Beetle Study

1) We build hypotheses for the situation

$$H_0\colon\ \mu = 0.49,$$
$$H_A\colon\ \mu < 0.49.$$

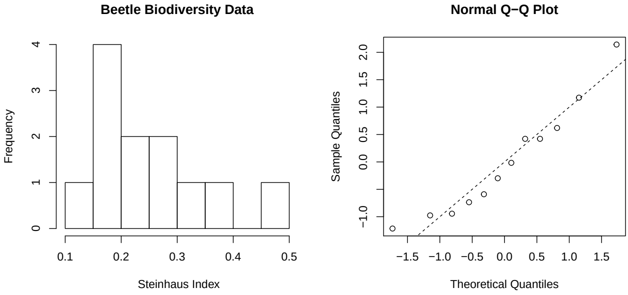
2) We want to build a hypothesis test using the t-distribution, so we have to check the normality of our population distribution.

Beetle Study

1) We build hypotheses for the situation

$$H_0\colon\ \mu = 0.49,$$
$$H_A\colon\ \mu < 0.49.$$

2) We want to build a hypothesis test using the t-distribution, so we have to check the normality of our population distribution.



Beetle Study

3) t-score your data

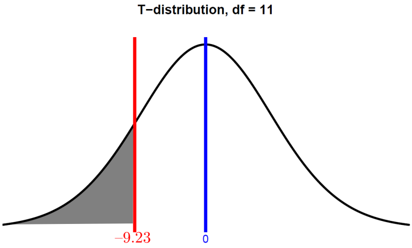
$$T = \frac{\text{point estimate} - \text{ null value}}{SE}$$
$$= \frac{0.25 - 0.49}{0.09/\sqrt{12}} \simeq -9.23.$$

Beetle Study

3) t-score your data

$$T = \frac{\text{point estimate} - \text{ null value}}{SE}$$
$$= \frac{0.25 - 0.49}{0.09/\sqrt{12}} \simeq -9.23.$$

4) Compute the p-value

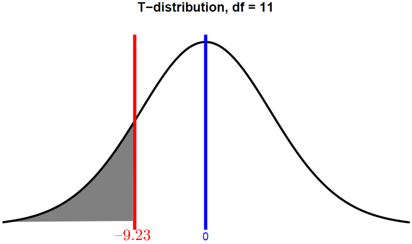


Beetle Study

3) t-score your data

$$T = \frac{\text{point estimate} - \text{ null value}}{SE}$$
$$= \frac{0.25 - 0.49}{0.09/\sqrt{12}} \simeq -9.23.$$

4) Compute the p-value



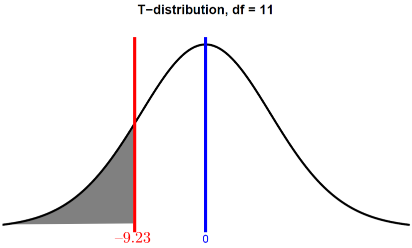
With Minitab, you find the area under the curve is $\simeq 8 \times 10^{-7}$.

Beetle Study

3) t-score your data

$$T = \frac{\text{point estimate} - \text{ null value}}{SE}$$
$$= \frac{0.25 - 0.49}{0.09/\sqrt{12}} \simeq -9.23.$$

4) Compute the p-value



With Minitab, you find the area under the curve is $\simeq 8 \times 10^{-7}$.

Since $p \simeq 8 \cdot 10^{-7} \ll 0.05$, we reject H_0 and favor H_A.

There is (a very) strong evidence that the true average biodiversity index is smaller than 0.49.

What Can Go Wrong?

There are Four possible scenarios we must think about related to hypothesis testing:

- We reject H_0 when H_0 is actually true
- We do not reject H_0 when H_0 is actually true
- We reject H_0 when H_0 is actually false
- We do not reject H_0 when H_0 is actually false

What Can Go Wrong?

There are Four possible scenarios we must think about related to hypothesis testing:

- We reject H_0 when H_0 is actually true Type I Error
- We do not reject H_0 when H_0 is actually true
- We reject H_0 when H_0 is actually false
- We do not reject H_0 when H_0 is actually false

What Can Go Wrong?

There are Four possible scenarios we must think about related to hypothesis testing:

- We reject H_0 when H_0 is actually true — Type I Error
- We do not reject H_0 when H_0 is actually true — Awesome!
- We reject H_0 when H_0 is actually false
- We do not reject H_0 when H_0 is actually false

What Can Go Wrong?

There are Four possible scenarios we must think about related to hypothesis testing:

- We reject H_0 when H_0 is actually true — Type I Error
- We do not reject H_0 when H_0 is actually true — Awesome!
- We reject H_0 when H_0 is actually false — Awesome!
- We do not reject H_0 when H_0 is actually false

What Can Go Wrong?

There are Four possible scenarios we must think about related to hypothesis testing:

- We reject H_0 when H_0 is actually true — Type I Error
- We do not reject H_0 when H_0 is actually true — Awesome!
- We reject H_0 when H_0 is actually false — Awesome!
- We do not reject H_0 when H_0 is actually false — Type II Error

What Can Go Wrong?

There are Four possible scenarios we must think about related to hypothesis testing:

- We reject H_0 when H_0 is actually true — Type I Error
- We do not reject H_0 when H_0 is actually true — Awesome!
- We reject H_0 when H_0 is actually false — Awesome!
- We do not reject H_0 when H_0 is actually false — Type II Error

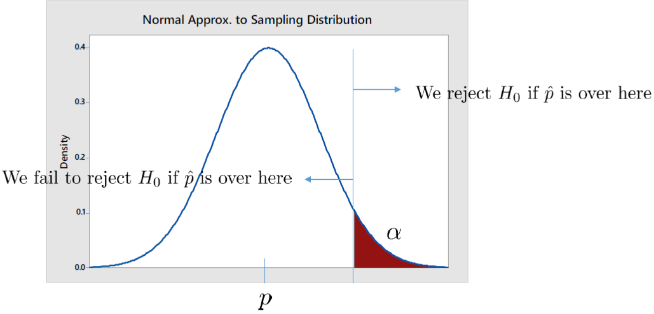
Truth \ Test Conclusion	Do not reject H_0	Reject H_0
H_0 true	OK	Type I Error
H_A true	Type II Error	OK

How Often Do We Make These Errors?

$$P(\text{Type I error}) = P(\text{reject } H_0 | H_0 \text{ is true}).$$

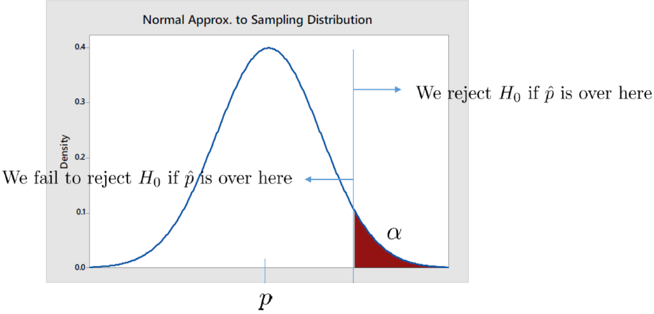
How Often Do We Make These Errors?

$$P(\text{Type I error}) = P(\text{reject } H_0 | H_0 \text{ is true}).$$



How Often Do We Make These Errors?

$$P(\text{Type I error}) = P(\text{reject } H_0 | H_0 \text{ is true}).$$



$$P(\text{Type I error}) = \alpha$$

We reject the null when the p-value is $< \alpha$ (say, 0.05). For this to happen, $\hat{p}$ must be large enough to get the shaded area $< \alpha$. Said differently, we need $\hat{p}$ to be in the top 5% of the data values. This happens with probability α.

How Often Do We Make These Errors?

We also care about

$$P\left(\text{Making a Type II Error}\right) = P\left(\text{Do not reject } H_0 | H_0 \text{ is false}\right).$$

How Often Do We Make These Errors?

We also care about

$$P\left(\text{Making a Type II Error}\right) = P\left(\text{Do not reject } H_0 | H_0 \text{ is false}\right).$$

No nice formula for this probability: it is given the letter β.
(Not related with to linear regression)

How Often Do We Make These Errors?

We also care about

$$P\,(\text{Making a Type II Error}) = P\,(\text{Do not reject } H_0 | H_0 \text{ is false})\,.$$

No nice formula for this probability: it is given the letter β.
(Not related with to linear regression)

The **power** of a test is

$$P\,(\text{Reject } H_0 | H_0 \text{ is false}) = 1 - \beta.$$

You'll learn how to deal with β in more advanced stat classes.

Type I and II Errors in Real Life

In the US legal system, you are either innocent (H_0) or guilty (H_A).

Type I and II Errors in Real Life

In the US legal system, you are either innocent (H_0) or guilty (H_A).

Type I Error: You reject H_0 when, in fact, H_0 is true.

Type I and II Errors in Real Life

In the US legal system, you are either innocent (H_0) or guilty (H_A).

Type I Error: You reject H_0 when, in fact, H_0 is true.
This means you convict an innocent person: "Wrongful conviction"

Type I and II Errors in Real Life

In the US legal system, you are either innocent (H_0) or guilty (H_A).

Type I Error: You reject H_0 when, in fact, H_0 is true.
This means you convict an innocent person: “Wrongful conviction”

Type II Error: You don’t reject H_0 when, in fact, H_0 is false.

Type I and II Errors in Real Life

In the US legal system, you are either innocent (H_0) or guilty (H_A).

Type I Error: You reject H_0 when, in fact, H_0 is true.
This means you convict an innocent person: “Wrongful conviction”

Type II Error: You don’t reject H_0 when, in fact, H_0 is false.
This means you set free a guilty person: “Guilty person set free”

Type I and II Errors in Real Life

In the US legal system, you are either innocent (H_0) or guilty (H_A).

Type I Error: You reject H_0 when, in fact, H_0 is true.
This means you convict an innocent person: "Wrongful conviction"

How you might reduce Type I Error: Require more proof to convict.
Notice that this increases Type II Error.

Type II Error: You don't reject H_0 when, in fact, H_0 is false.
This means you set free a guilty person: "Guilty person set free"

How you might reduce Type II Error: Require less proof to convict.
Notice that this increases Type I Error.

Type I and II Errors in Real Life

In the US legal system, you are either innocent (H_0) or guilty (H_A).

Type I Error: You reject H_0 when, in fact, H_0 is true.
This means you convict an innocent person: "Wrongful conviction"

How you might reduce Type I Error: Require more proof to convict.
Notice that this increases Type II Error.

Type II Error: You don't reject H_0 when, in fact, H_0 is false.
This means you set free a guilty person: "Guilty person set free"

How you might reduce Type II Error: Require less proof to convict.
Notice that this increases Type I Error.

Moral: For most actions,

$$\text{Type I Error} \searrow \text{when Type II Error} \nearrow \text{(and vice versa).}$$

A food inspector is asked to investigate a restaurant that has some negative sanitation reputation. The inspector decides to use a hypothesis testing framework to decide if regulations are being met. If the restaurant is in violation of these regulations, the restaurant will be closed down.

A food inspector is asked to investigate a restaurant that has some negative sanitation reputation. The inspector decides to use a hypothesis testing framework to decide if regulations are being met. If the restaurant is in violation of these regulations, the restaurant will be closed down.

What are the null and alternative hypotheses?

A food inspector is asked to investigate a restaurant that has some negative sanitation reputation. The inspector decides to use a hypothesis testing framework to decide if regulations are being met. If the restaurant is in violation of these regulations, the restaurant will be closed down.

What are the null and alternative hypotheses?

H_0: The restaurant meets the regulations
H_A: The restaurant violates sanitation regulations

A food inspector is asked to investigate a restaurant that has some negative sanitation reputation. The inspector decides to use a hypothesis testing framework to decide if regulations are being met. If the restaurant is in violation of these regulations, the restaurant will be closed down.

What are the null and alternative hypotheses?

H_0: The restaurant meets the regulations
H_A: The restaurant violates sanitation regulations

What does Type I Error mean in this case?

A food inspector is asked to investigate a restaurant that has some negative sanitation reputation. The inspector decides to use a hypothesis testing framework to decide if regulations are being met. If the restaurant is in violation of these regulations, the restaurant will be closed down.

What are the null and alternative hypotheses?

H_0: The restaurant meets the regulations
H_A: The restaurant violates sanitation regulations

What does Type I Error mean in this case?

A Type I Error would mean the restaurant did meet regulations, but the inspector thought this was not true and shut it down.

What does Type II Error mean?

What does Type II Error mean?

A Type II Error would mean the restaurant violated regulations, but the inspector made the mistake of letting the restaurant stay open.

What does Type II Error mean?

A Type II Error would mean the restaurant violated regulations, but the inspector made the mistake of letting the restaurant stay open.

Which error is worse for the restaurant owner? Which is worse for the diners?

What does Type II Error mean?

A Type II Error would mean the restaurant violated regulations, but the inspector made the mistake of letting the restaurant stay open.

Which error is worse for the restaurant owner? Which is worse for the diners?

The restaurant owner dislikes Type I because the restaurant is shut down for no reason.
Diners dislike Type II because they are exposed to dangerous conditions.

Reviewing the Framework: 21 Questions

A student claims she can guess the suit on the top card from a shuffled deck better than chance. You decide to design an experiment to test this.

1. We will study the parameter p. What should p represent?

Reviewing the Framework: 21 Questions

A student claims she can guess the suit on the top card from a shuffled deck better than chance. You decide to design an experiment to test this.

1. We will study the parameter p. What should p represent?

Let p be the proportion of suits the student can correctly name using her special powers (out of all possible cards we could ever ask her about).

Reviewing the Framework: 21 Questions

A student claims she can guess the suit on the top card from a shuffled deck better than chance. You decide to design an experiment to test this.

1. We will study the parameter p. What should p represent?

Let p be the proportion of suits the student can correctly name using her special powers (out of all possible cards we could ever ask her about).

2. What is an appropriate null hypothesis?

Reviewing the Framework: 21 Questions

A student claims she can guess the suit on the top card from a shuffled deck better than chance. You decide to design an experiment to test this.

1. We will study the parameter p. What should p represent?

Let p be the proportion of suits the student can correctly name using her special powers (out of all possible cards we could ever ask her about).

2. What is an appropriate null hypothesis?

If the student has no powers, she will get the correct answer one-fourth of the time. We set H_0 $p = 0.25$.

3. What is our alternative hypothesis?

3. What is our alternative hypothesis?

Since she claims to do better, we have H_A: $p > 0.25$.

3. What is our alternative hypothesis?

Since she claims to do better, we have H_A: $p > 0.25$.

4. When we draw a sample, how do we make sure the suits of the cards are independent of one another?

3. What is our alternative hypothesis?

Since she claims to do better, we have H_A: $p > 0.25$.

4. When we draw a sample, how do we make sure the suits of the cards are independent of one another?

We need to shuffle the cards after each top card is guessed and revealed. Otherwise, the student has information about the suits that remain in the deck.

3. What is our alternative hypothesis?

Since she claims to do better, we have H_A: $p > 0.25$.

4. When we draw a sample, how do we make sure the suits of the cards are independent of one another?

We need to shuffle the cards after each top card is guessed and revealed. Otherwise, the student has information about the suits that remain in the deck.

5. How many cards should we test the student on and why?

3. What is our alternative hypothesis?

Since she claims to do better, we have H_A: $p > 0.25$.

4. When we draw a sample, how do we make sure the suits of the cards are independent of one another?

We need to shuffle the cards after each top card is guessed and revealed. Otherwise, the student has information about the suits that remain in the deck.

5. How many cards should we test the student on and why?

We eventually look at a sampling distribution, which will be normal if we have at least 10 successes and failures. We expect $0.25 \times n$ successes, so we need at least $n \geq 40$ to get adequate successes.

6. We plan to also calculate a confidence interval for p based on our sample. Name two ways to get a smaller interval.

6. We plan to also calculate a confidence interval for p based on our sample. Name two ways to get a smaller interval.

The CI is $\hat{p} \pm z^* SE(\hat{p})$. If we decrease the confidence level or increase the sample size, this will shrink the interval.

6. We plan to also calculate a confidence interval for p based on our sample. Name two ways to get a smaller interval.

The CI is $\hat{p} \pm z^* SE(\hat{p})$. If we decrease the confidence level or increase the sample size, this will shrink the interval.

7. We decide on a sample of size 100 because $100 \geq 40$ and drawing a larger sample isn't too hard. During the experiment, she guesses 32 suits correctly. She says this proves her claim. Why is her logic false?

6. We plan to also calculate a confidence interval for p based on our sample. Name two ways to get a smaller interval.

The CI is $\hat{p} \pm z^* SE(\hat{p})$. If we decrease the confidence level or increase the sample size, this will shrink the interval.

7. We decide on a sample of size 100 because $100 \geq 40$ and drawing a larger sample isn't too hard. During the experiment, she guesses 32 suits correctly. She says this proves her claim. Why is her logic false?

We got $\hat{p} = 0.32$, but sampling variability exists. Some samples of 100 cards would yield results like this, other may not support her claim. We must sort out whether 32% is from sampling variation or not.

8. If we drew many samples of size 100, found the proportion of correct guesses of each and plotted these, we would get a sampling distribution (histogram). Why is it approximately normal in this case?

8. If we drew many samples of size 100, found the proportion of correct guesses of each and plotted these, we would get a sampling distribution (histogram). Why is it approximately normal in this case?

Reshuffling gives independence, 100 cards i less than 10% of all the cards we could ask her ($= \infty$!), and the sample is big enough to expect at least 10 successes and failures.

8. If we drew many samples of size 100, found the proportion of correct guesses of each and plotted these, we would get a sampling distribution (histogram). Why is it approximately normal in this case?

Reshuffling gives independence, 100 cards i less than 10% of all the cards we could ask her ($= \infty$!), and the sample is big enough to expect at least 10 successes and failures.

9. Where is the normal approximation to the sampling distribution centered? What spread does it have?

8. If we drew many samples of size 100, found the proportion of correct guesses of each and plotted these, we would get a sampling distribution (histogram). Why is it approximately normal in this case?

Reshuffling gives independence, 100 cards i less than 10% of all the cards we could ask her (= ∞!), and the sample is big enough to expect at least 10 successes and failures.

9. Where is the normal approximation to the sampling distribution centered? What spread does it have?

Under the null,
$\mu(\hat{p}) = p = 0.25$ and $SE(\hat{p}) = \sqrt{pq/n} = \sqrt{0.75 \times 0.25/100} \simeq 0.0433$.

8. If we drew many samples of size 100, found the proportion of correct guesses of each and plotted these, we would get a sampling distribution (histogram). Why is it approximately normal in this case?

Reshuffling gives independence, 100 cards i less than 10% of all the cards we could ask her ($= \infty!$), and the sample is big enough to expect at least 10 successes and failures.

9. Where is the normal approximation to the sampling distribution centered? What spread does it have?

Under the null,
$\mu(\hat{p}) = p = 0.25$ and $SE(\hat{p}) = \sqrt{pq/n} = \sqrt{0.75 \times 0.25/100} \simeq 0.0433.$

10. In words, what will be the p-value for this situation represent?

8. If we drew many samples of size 100, found the proportion of correct guesses of each and plotted these, we would get a sampling distribution (histogram). Why is it approximately normal in this case?

Reshuffling gives independence, 100 cards i less than 10% of all the cards we could ask her (= ∞!), and the sample is big enough to expect at least 10 successes and failures.

9. Where is the normal approximation to the sampling distribution centered? What spread does it have?

Under the null,
$\mu(\hat{p}) = p = 0.25$ and $SE(\hat{p}) = \sqrt{pq/n} = \sqrt{0.75 \times 0.25/100} \simeq 0.0433$.

10. In words, what will be the p-value for this situation represent?

It is the probability that sample variation gives us the result $\hat{p} = 0.32$ (or something more extreme) if the null hypothesis is true (that is, if the student is just randomly guessing).

11. Using symbols, how can we fill in this blank for our particular problem?

$$p\text{-value} = P(\qquad\qquad\qquad\qquad)$$

11. Using symbols, how can we fill in this blank for our particular problem?

$$p\text{-value} = P(\hat{p} \geq 0.32 | p = 0.25)$$

12. Describe two ways to find this p-value.

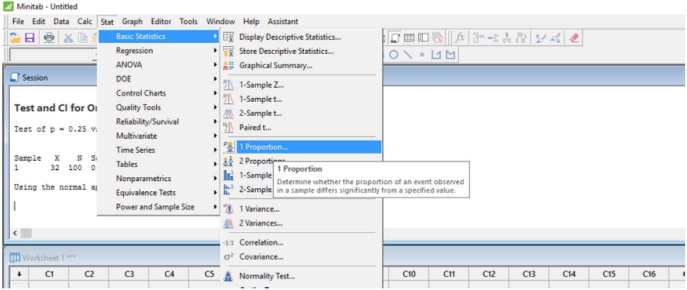
11. Using symbols, how can we fill in this blank for our particular problem?

$$p\text{-value} = P(\hat{p} \geq 0.32 | p = 0.25)$$

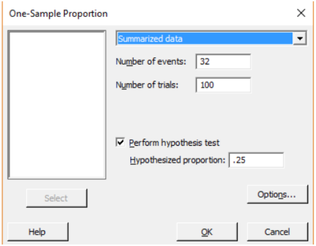
12. Describe two ways to find this p-value.

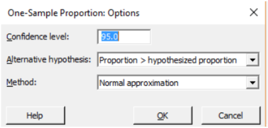
- We can use Minitab to find the area on the sampling distribution itself,
- or we can convert $\hat{p}$ to a z-score and use a table to find the area on the standard normal distribution.

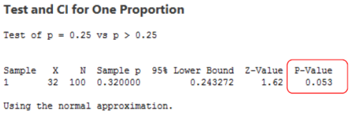
Let's learn a new way to do this!



Minitab allows us to run tests and gives P-values (without drawing areas or showing any of the visual thinking behind the test).







13. Our p-value was 0.053. Interpret what this means in terms of probability in our situation.

13. Our p-value was 0.053. Interpret what this means in terms of probability in our situation.

If the student is randomly guessing, we would expect to get a result like 0.32 correct (or more) about 5.3% of the time.

13. Our p-value was 0.053. Interpret what this means in terms of probability in our situation.

If the student is randomly guessing, we would expect to get a result like 0.32 correct (or more) about 5.3% of the time.

14. This suggests that if we ran the experiment 20 times, one of those times we would expect to get a result like 0.32 (or wilder!). If there were 20 science labs in the country running this experiment, what would happen?

13. Our p-value was 0.053. Interpret what this means in terms of probability in our situation.

If the student is randomly guessing, we would expect to get a result like 0.32 correct (or more) about 5.3% of the time.

14. This suggests that if we ran the experiment 20 times, one of those times we would expect to get a result like 0.32 (or wilder!). If there were 20 science labs in the country running this experiment, what would happen?

We expect one of them to get a result whose probability is very near the 0.05 significance threshold. They will probably publish a paper saying they found a student with special powers.
If we live in a society that only publishes results with p-values < 0.05, then for every published finding, we might expect 19 unpublished findings that found no interesting result.

15. The p-value is above the 0.05 cutoff. What do many scientists do?

15. The p-value is above the 0.05 cutoff. What do many scientists do?

They fail to reject the null hypothesis. The 0.32 evidence is not enough to convince us the girl has special powers. We keep believing she is randomly guessing.

15. The p-value is above the 0.05 cutoff. What do many scientists do?

They fail to reject the null hypothesis. The 0.32 evidence is not enough to convince us the girl has special powers. We keep believing she is randomly guessing.

16. Give some reasons why a 0.05 cutoff is silly.

15. The p-value is above the 0.05 cutoff. What do many scientists do?

They fail to reject the null hypothesis. The 0.32 evidence is not enough to convince us the girl has special powers. We keep believing she is randomly guessing.

16. Give some reasons why a 0.05 cutoff is silly.

0.05 is a historical artifact derived from a 1931 book by Fisher (*The Design of Experiments*). He though a 1 in 20 event might be surprising enough to toss out one's belief system (H_0).
In some situations, you might be willing to reject the null hypothesis more or less readily.

Are the Following True or False? Why?

17. The probability that the null hypothesis is true is 5.3%.

Are the Following True or False? Why?

17. The probability that the null hypothesis is true is 5.3%.

False. We found $P(\hat{p} \geq 0.32 | p = 0.25) = 0.053$, **not** "$P(H_0$ is true$) = 0.053$"

Are the Following True or False? Why?

17. The probability that the null hypothesis is true is 5.3%.

False. We found $P(\hat{p} \geq 0.32 | p = 0.25) = 0.053$, **not** "$P(H_0$ is true$) = 0.053$"

18. If we reject the null hypothesis, this means the alternative hypothesis is true.

Are the Following True or False? Why?

17. The probability that the null hypothesis is true is 5.3%.

False. We found $P(\hat{p} \geq 0.32 | p = 0.25) = 0.053$, **not** "$P(H_0$ is true$) = 0.053$"

18. If we reject the null hypothesis, this means the alternative hypothesis is true.

False. If we reject H_0, it is because we had evidence **against** H_0 and for H_A. This doesn't make H_A true.

Are the Following True or False? Why?

17. The probability that the null hypothesis is true is 5.3%.

False. We found $P(\hat{p} \geq 0.32 | p = 0.25) = 0.053$, **not** "$P(H_0 \text{ is true}) = 0.053$"

18. If we reject the null hypothesis, this means the alternative hypothesis is true.

False. If we reject H_0, it is because we had evidence **against** H_0 and for H_A. This doesn't make H_A true.

19. The choice of the cutoff is influenced by the sample size n.

Are the Following True or False? Why?

17. The probability that the null hypothesis is true is 5.3%.

False. We found $P(\hat{p} \geq 0.32 | p = 0.25) = 0.053$, **not** "$P(H_0$ is true$) = 0.053$"

18. If we reject the null hypothesis, this means the alternative hypothesis is true.

False. If we reject H_0, it is because we had evidence **against** H_0 and for H_A. This doesn't make H_A true.

19. The choice of the cutoff is influenced by the sample size n.

False. The cutoff, or α level, is decided by the researcher. Values of $\alpha = 0.05, 0.01, 0.001$ are common.

Are the Following True or False? Why?

20. It is better to use $\alpha = 0.01$ than $\alpha = 0.05$.

Are the Following True or False? Why?

20. It is better to use $\alpha = 0.01$ than $\alpha = 0.05$.

Not necessarily, this just manipulates the Type I Error rate.

Are the Following True or False? Why?

20. It is better to use $\alpha = 0.01$ than $\alpha = 0.05$.

Not necessarily, this just manipulates the Type I Error rate.

21. If $\alpha = 0.10$, the the p-value of 0.09 is statistically significant.

Are the Following True or False? Why?

20. It is better to use $\alpha = 0.01$ than $\alpha = 0.05$.

Not necessarily, this just manipulates the Type I Error rate.

21. If $\alpha = 0.10$, the the p-value of 0.09 is statistically significant.

True. The phrase "statistically significant" means that the p-value is less than the α cutoff.

Note: Statistical significance is not the same as **practical significance**. For example, consider an SAT class that raises grades by 0.1% vs. a drug that cures 0.1% more people.

One Final Example: The Case of Two Coins

Your friend has two coins, one that is fair, and the other that comes up tails 85% of the time. You get to spin one of the coins and then must decide which coin it is.

One Final Example: The Case of Two Coins

Your friend has two coins, one that is fair, and the other that comes up tails 85% of the time. You get to spin one of the coins and then must decide which coin it is.

1. You pick a coin and flip it under the assumption it is unfair. What are H_0 and H_A? Let p be the probability of getting tails on the flipped coin.

One Final Example: The Case of Two Coins

Your friend has two coins, one that is fair, and the other that comes up tails 85% of the time. You get to spin one of the coins and then must decide which coin it is.

1. You pick a coin and flip it under the assumption it is unfair. What are H_0 and H_A? Let p be the probability of getting tails on the flipped coin.

Set H_0: $p = 0.85$ and H_A: $p = 0.5$.

One Final Example: The Case of Two Coins

Your friend has two coins, one that is fair, and the other that comes up tails 85% of the time. You get to spin one of the coins and then must decide which coin it is.

1. You pick a coin and flip it under the assumption it is unfair. What are H_0 and H_A? Let p be the probability of getting tails on the flipped coin.

Set H_0: $p = 0.85$ and H_A: $p = 0.5$.

2. Your flip is head. What would you do?

One Final Example: The Case of Two Coins

Your friend has two coins, one that is fair, and the other that comes up tails 85% of the time. You get to spin one of the coins and then must decide which coin it is.

1. You pick a coin and flip it under the assumption it is unfair. What are H_0 and H_A? Let p be the probability of getting tails on the flipped coin.

Set H_0: $p = 0.85$ and H_A: $p = 0.5$.

2. Your flip is head. What would you do?

Find a p-value! It equals

$$P(\text{our result or something more extreme}|H_0 \text{ is true}).$$

Here, this means $P(\text{Heads}|H_0) = 0.15$. We might reject H_0.

3. If your flip was tails, what do you do?

3. If your flip was tails, what do you do?

Here, the p-value is $P(\text{Tails}|H_0) = 0.0.85$. We have a piece of information that meshes well with H_0, so we probably won't reject it.

3. If your flip was tails, what do you do?

Here, the p-value is $P(\text{Tails}|H_0) = 0.0.85$. We have a piece of information that meshes well with H_0, so we probably won't reject it.

4. What do Type I and Type II Errors mean in context?

3. If your flip was tails, what do you do?

Here, the p-value is $P(\text{Tails}|H_0) = 0.0.85$. We have a piece of information that meshes well with H_0, so we probably won't reject it.

4. What do Type I and Type II Errors mean in context?

	Heads Appears	Tails Appears
H_0 is true (unfair coin was flipped)	We think our coin is FAIR Type I Error	We think our coin is UNFAIR
H_0 is false (fair coin was flipped)	We think our coin is FAIR.	We think our coin is UNFAIR. Type 2 Error