Chapter 8

Inference about a Population Mean
One-sample t-procedures

The power of a test

The **power** of a test of hypothesis with fixed significance level α is the probability that the test will reject the null hypothesis when a particular alternative value of the parameter is true.

In other words, power is the probability that the data gathered in an experiment will be sufficient to reject a wrong null hypothesis.

Knowing the power of your test is important:

- When designing your experiment: select a sample size large enough to detect an effect of a magnitude you think is meaningful.
- When a test found no significance: Check that your test would have had enough power to detect an effect of a magnitude you think is meaningful.

Planning studies: the power of a statistical test

- overview of influences on “How many observations do I need?”
  - If you insist on a smaller significance level (such as 1% rather than 5%), you have to take a larger sample. **A smaller significance level requires stronger evidence to reject the null hypothesis**.
  - If you insist on **higher power** (such as 99% rather than 90%), you will need a **larger sample**. **Higher power gives a better chance of detecting a difference when it is really there**.
  - At any significance level and desired power, **a two-sided alternative requires a larger sample than a one-sided alternative**.
  - At any significance level and desired power, **detecting a small effect requires a larger sample than detecting a large effect**.

Type I and II errors

- A **Type I error** is made when we reject the null hypothesis and the null hypothesis is actually true (incorrectly reject a true H_0).

  The probability of making a Type I error is the significance level α.

- A **Type II error** is made when we fail to reject the null hypothesis and the null hypothesis is false (incorrectly keep a false H_0).

  The probability of making a Type II error is labeled β.

  The power of a test is $1 - \beta$.

Running a test of significance is a balancing act between the chance α of making a **Type I error** and the chance β of making a **Type II error.** Reducing α reduces the power of a test and thus increases β.

	H_0 true	H_a true
Reject H_0	Type I error	Correct decision
Accept H_0	Correct decision	Type II error

It might be tempting to emphasize greater power (the more the better).

- However, with "too much power" trivial effects become highly significant.
- A type II error is not definitive since a failure to reject the null hypothesis does not imply that the null hypothesis is wrong.

One sample t-procedure objectives

- Conditions for inference about a mean
- The *t* distributions
- The one-sample *t* confidence interval
- The one-sample *t* test
- Matched pairs *t* procedures
- Robustness of *t* procedures

Conditions for inference about a mean

- We can regard our data as a **simple random sample** (SRS) from the population. This condition is very important.

- Observations from the population have a **Normal distribution** with mean μ and standard deviation σ. In practice, it is enough that the distribution be symmetric and single-peaked, unless the sample is very small.

- Both μ and σ are unknown parameters.

Conditions for inference about a mean

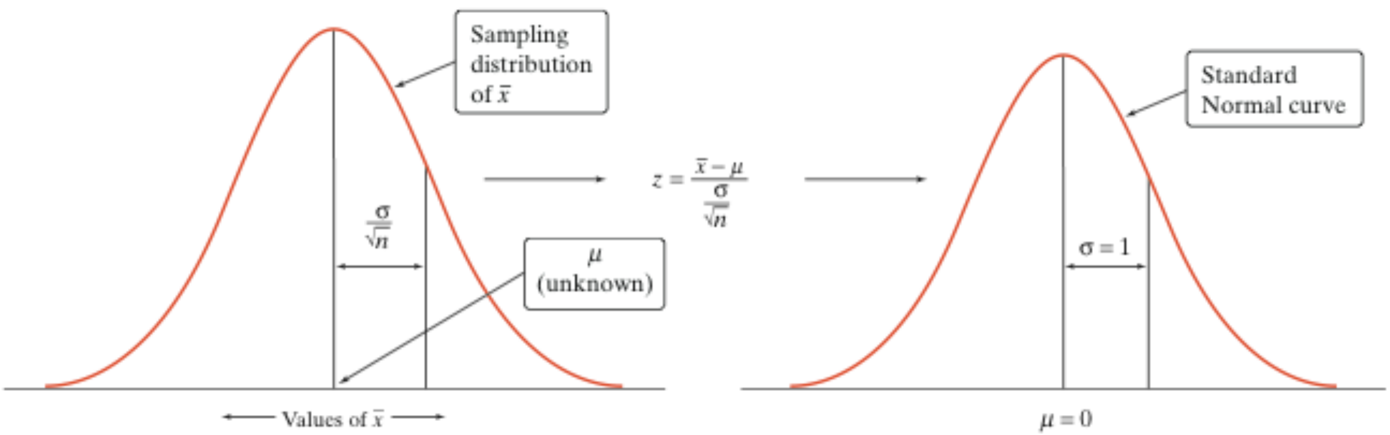
- First, a caution: A condition that applies to all the inference methods in this book**. *The population must be much larger than the sample, say at least 20 times as large*.**
- Previously, we assumed that we knew the population standard deviation, σ, and thus were able to compute probabilities for a sample mean. In practice, we do not know σ. Because we don't know σ, we estimate it by the sample standard deviation *s*. We then estimate the standard deviation of $\bar{x}$ by $s/\sqrt{n}$.
- This quantity $s/\sqrt{n}$ is called the ***standard error*** of the sample mean $\bar{x}$.

STANDARD ERROR

- **When the standard deviation of a statistic is estimated from data, the result is called the standard error of the statistic. The standard error of the sample mean $\bar{x}$ is $s/\sqrt{n}$.**

The t distributions

- When the sampling distribution of $\bar{x}$ is close to Normal and σ is known, we can find probabilities involving $\bar{x}$ by standardizing: $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$.



- When we don't know σ, we estimate it by the sample standard deviation *s*. What happens when we standardize?

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

- **This new statistic does not have a Normal distribution.**

The *t* distributions

- When we standardize based on the sample standard deviation *s*, our statistic has a new distribution called a *t* distribution.

THE ONE-SAMPLE *t* STATISTIC AND THE *t* DISTRIBUTIONS

- Draw an SRS of size *n* from a large population that has the Normal distribution with mean μ and standard deviation σ. The one-sample *t* statistic

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

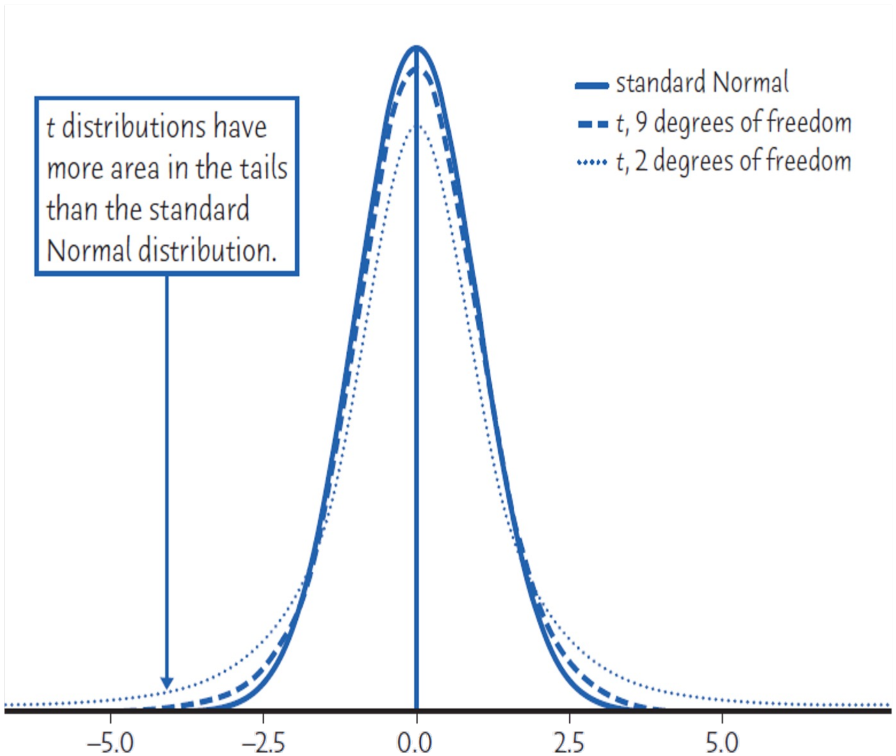
has the *t* distribution with *n* – 1 degrees of freedom.

- The *t* statistic has the same interpretation as any standardized statistic: it says how far $\bar{x}$ is from its mean μ in standard deviation units.

The *t* distributions

- There is a different *t* distribution for each sample size, specified by its **degrees of freedom**.
- When we perform inference about a population mean *μ* using a *t* distribution, the appropriate degrees of freedom are found by subtracting 1 from the sample size *n*, making $df = n - 1$. We will write the *t* distribution with *n* – 1 degrees of freedom as t_{n-1}.
- It has a different shape than the standard Normal curve:
  - It is symmetric with a single peak at 0.
  - It has much more area in the tails.

The *t* distributions



t distributions and standard Normal distribution

- The density curves of the *t* distributions are similar in shape to the standard Normal curve. They are symmetric about 0, single peaked, and bell shaped.
- The variability of the *t* distributions is a bit greater than that of the standard Normal distribution. The *t* distributions have more probability in the tails and less in the center than does the standard Normal. This is true because substituting the estimate *s* for the fixed parameter σ introduces more variation into the statistic.
- the degrees of freedom increase, the *t* density curve approaches the $N(0, 1)$ curve ever more closely. This happens because *s* estimates σ more accurately as the sample size increases. So using s in place of σ causes little extra variation when the sample is large.

Using the *t* Procedures

- The *t*-procedures (CI and test) are exactly correct when the distribution of the population is exactly normal.
- The usefulness of the *t* procedures depends on how strongly they are affected by lack of Normality.
- **Robust:** if the confidence level or *P*-value does not change very much when the conditions for use of the procedure are violated.
- The assumption that the data are an SRS.
- ***Sample size less than 15:*** use it if the data is close to Normal (symmetric, single peak, no outliers).
- ***Sample size at least 15:*** except: outliers or strong skewness
- ***Large samples:*** use it when $n \geq 40$.

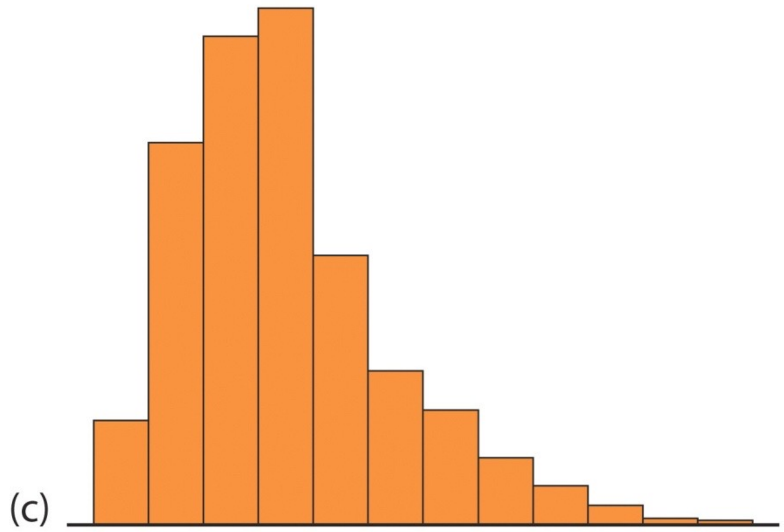
Can we use *t*?

(b)

23	0
24	0
25	
26	5
27	
28	7
29	
30	259
31	399
32	033677
33	0236

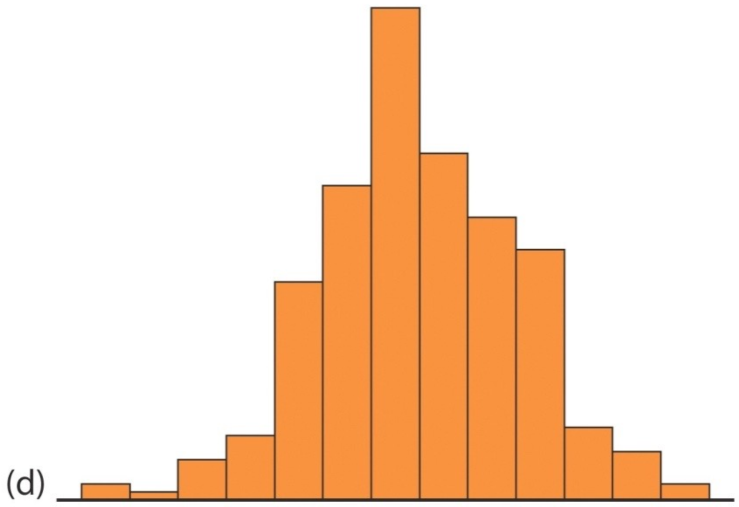
- This stemplot shows the force required to pull apart 20 pieces of Douglas fir.
- **Cannot use *t*.** The data are strongly skewed to the left, so we cannot trust the *t* procedures for $n = 20$.

Can we use *t*?



- This histogram shows the distribution of word lengths in Shakespeare's plays (n=200).
- **Can use *t*.** The data is skewed right, but there are no outliers. We can use the *t* procedures since $n \geq 40$.

Can we use *t*?



- This histogram shows the heights of college students.
- **Can use *t*.** The distribution is close to Normal, so we can trust the *t* procedures for any sample size.

One-sample *t* confidence interval

- The *one-sample* t *interval for a population mean* is similar in both reasoning and computational detail to the one-sample *z* interval for a population proportion.

THE ONE-SAMPLE *t* CONFIDENCE INTERVAL

- Draw an SRS of size *n* from a large population with unknown mean μ. A level *C* confidence interval for μ is:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the critical value for the t_{n-1} density curve with area *C* between $-t^*$ and t^*. This interval is exact when the population distribution is Normal and is approximately correct for large n in other cases.

Using Table C

Suppose you want to construct a 95% confidence interval for the mean μ of a Normal population based on an SRS of size n = 12. What critical t^* should you use?

	Confidence level			
df	***90%***	95%	96%	98%
10	1.812	2.228	2.359	2.764
11	1.796	2.201	2.328	2.718
12	1.782	2.179	2.303	2.681
z*	1.645	1.960	2.054	2.326
1-sided p	0.05	0.025	0.02	0.01
2-sided p	0.10	0.05	0.04	0.02

In Table C, we consult the row corresponding to df = n – 1 = 11.

We move across that row to the entry that is directly above the 95% confidence level.

The desired critical value is t^* = 2.201.

Example

- STATE: Does the expectation of good weather lead to more generous behavior? Psychologists studied the size of the tip in a restaurant when a message indicating that the next day's weather would be good was written on the bill. Here are tips from 20 patrons, measured in percent of the total bill:

20.8	18.7	19.9	20.6	21.9	23.4	22.8	24.9	22.2	20.3
24.9	22.3	27.0	20.4	22.2	24.0	21.1	22.1	22.0	22.7

  We want to estimate the mean tip for comparison with tips under the other conditions.
- PLAN: We will estimate the mean percentage tip μ for all patrons of this restaurant when they receive a message on their bill indicating that the next day's weather will be good by giving a 95% confidence interval.
- SOLVE: Checking conditions:

  SRS: This study was actually a randomized comparative experiment; as a consequence, we are willing to regard these patrons as an SRS from all patrons of this restaurant.

Example

- **Normal distribution:** We can't look at the population, but we can examine the sample. The stemplot does not suggest any strong departures from Normality. The sample size is at least 15, so we are comfortable using the *t* interval.

18	7
19	9
20	3 4 6 8
21	1 9
22	0 1 2 2 3 7 8
23	4
24	0 9 9
25	
26	
27	0

- We can proceed to calculation. For these data, $\bar{x}$ = 22.21 and *s* = 1.963.

- The degrees of freedom are $n - 1$ = 19. From Table C, we find that, for 95% confidence, t^* = 2.093. The confidence interval is:

$$
\begin{aligned}
\bar{x} \pm t^* \frac{s}{\sqrt{n}} &= 22.21 \pm 2.093 \frac{1.963}{\sqrt{20}} \\
&= 22.21 \pm 0.92 \\
&= 21.29 \text{ to } 23.13 \text{ "percent"}
\end{aligned}
$$

- CONCLUDE: We are 95% confident that the mean percent tip for all patrons of this restaurant, when their bill contains a message that the next day's weather will be good, is between 21.29 and 23.13.

The one-sample *t*-test

A test of hypotheses requires a few steps:

1. Stating the null and alternative hypotheses (H_0 versus H_a)
2. Choosing a significance level α
3. Calculating t and its degrees of freedom
4. Finding the area under the curve *with Table C*
5. Stating the P-value and interpreting the result

The one-sample *t* test

Testing hypotheses about the mean μ of a Normal population follows the same reasoning as for testing hypotheses about a population proportion that we met earlier.

THE ONE-SAMPLE *t* TEST

- Draw an SRS of size n from a large population having unknown mean μ. To **test the hypothesis H_0: $\mu = \mu_0$**, compute the one-sample *t* statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- These *P*-values are exact if the population distribution is Normal and are approximately correct for large *n* in other cases.

In terms of a variable T having the t_{n-1} distribution, the ***P-value*** is the probability, when assuming H_0 is true, of randomly drawing a sample like the one obtained or more extreme, in the direction of H_a.

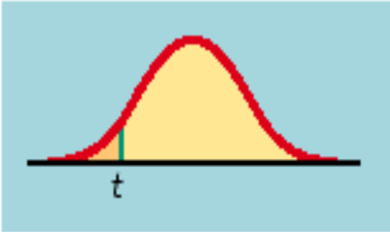
These *P*-values **are exact** if the population distribution **is Normal** and **are approximately correct** for **large *n*** in other cases.

The P-value is calculated as the corresponding area under the curve, one-tailed or two-tailed **depending on *H*a**:

One-sided (one-tailed)

$H_a\colon \mu > \mu_0 \Rightarrow P(T \geq t)$

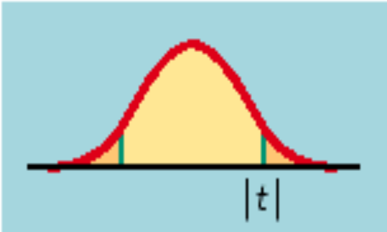
$H_a\colon \mu < \mu_0 \Rightarrow P(T \leq t)$



$$\boxed{t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}}$$

Two-sided (two-tailed)

$H_a\colon \mu \neq \mu_0 \Rightarrow 2P(T \geq |t|)$



One sample T-test: Case Study I

- Rosie is an aging sheep dog who gets regular checkups from her owner, the local vet. Let X be a random variable that represents Rosie's resting heart rate (in beats per minute). The vet checked the Merck Veterinary Manual and found that for dogs of this breed, the mean resting heart rate is 115 b/m.
- Over the past six weeks Rosie's heart rate (b/m) measured: 93, 109, 110, 89, 112, 117
- The sample mean is $\bar{x} = 105$ b/m. Sample s.d S=11.26 b/m
- The vet concerned that Rosie's heart rate may be slowing. Do the data indicate that this is the case?
- What is the standard error of $\bar{x}$?

Case Study I

- Are the data normal?

8|9

9|3

10|9

11|0 2 7

Hard to judge Normality from just 6 observations, a stemplot of the data shows no outliers, clusters, or extreme skewness. Thus, *P*-values for the *t* test will be reasonably accurate.

Case Study I

1. ***Hypotheses:*** H_0:
   H_a:
2. ***Test Statistic:***
   (df = ?)
   $$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = ?$$
3. ***P-value:***
   P-value = $P(T \leq -2.18) = 0.0405565$ *(using a computer)*
   P-value is between 0.025 and 0.05 since $t = |-2.18|$ is between $t^* = 2.015$ ($p = 0.05$) and $t^* = 2.571$ ($p = 0.025$) *(Table D)*
4. ***Conclusion:***
   Since the P-value is smaller than $\alpha = 0.05$, there is strong evidence that Rosie's heart rate is slowing down.

JMP example for one sample t-procedure

- 20.30 What do you think of Mitt?
- (a) Make a stem-plot. Is there any sign of major deviation from Normality?
- (b) Give a 95% confidence interval for the mean rating
- (c) Is there significant evidence at 5% level that the mean rating is greater than 3.5 (a neutral rating).

- Open data, click Analyze->Distribution, select Mitt Rating into Y box, click ok. Then change to horizontal layout.
- For 95% CI: Click the red triangle besides Mitt Rating, click Confidence interval->0.95. This is the t-procedure CI.
- For HT, click the red triangle besides Mitt Rating, click Test mean, enter the hypothesized mean 3.5, click ok. You will get the test-statistic. You need to look at your Ha to find the right p-value. Here for the right-sided test, p-value is 0.0032. Since $p<0.05$, we reject H0, and conclude that there is very strong evidence that the mean rating is greater than 3.5.

Matched Pairs *t* Procedures

- To compare two treatments, subjects are matched in pairs and each treatment is given to one subject in each pair.
- Before-and-after observations on the same.
- Apply the one-sample *t* procedures to the observed *differences*.
  $X_{\text{difference}} = (X_1 - X_2)$
- The parameter *m* is the mean difference
- H_0: $\mu_{\text{difference}} <= 0$ (or >=0, or =0) ; H_a: $\mu_{\text{difference}} > 0$ (or <0, or ≠0)
- **Conceptually, this is not different from tests on one sample**

Matched Pair Case Study

Sweetening Colas

Cola makers test new recipes for loss of sweetness during storage. Trained tasters rate the sweetness before and after storage. Here are the sweetness losses (sweetness before storage minus sweetness after storage) found by 10 tasters for a new cola recipe:

2.0 0.4 0.7 2.0 -0.4 2.2 -1.3 1.2 1.1 2.3

Are these data good evidence that the cola lost sweetness during storage?

Matched Pair Case Study

Sweetening Colas

Are these 10 carefully trained tasters an SRS?

Hard to judge Normality from just 10 observations, a stemplot of the data shows no outliers, clusters, or extreme skewness. Thus, *P*-values for the *t* test will be reasonably accurate.

−1	3
−0	4
0	47
1	12
2	0023

Matched Pair Case Study

1. ***Hypotheses:*** H_0:
   H_a:

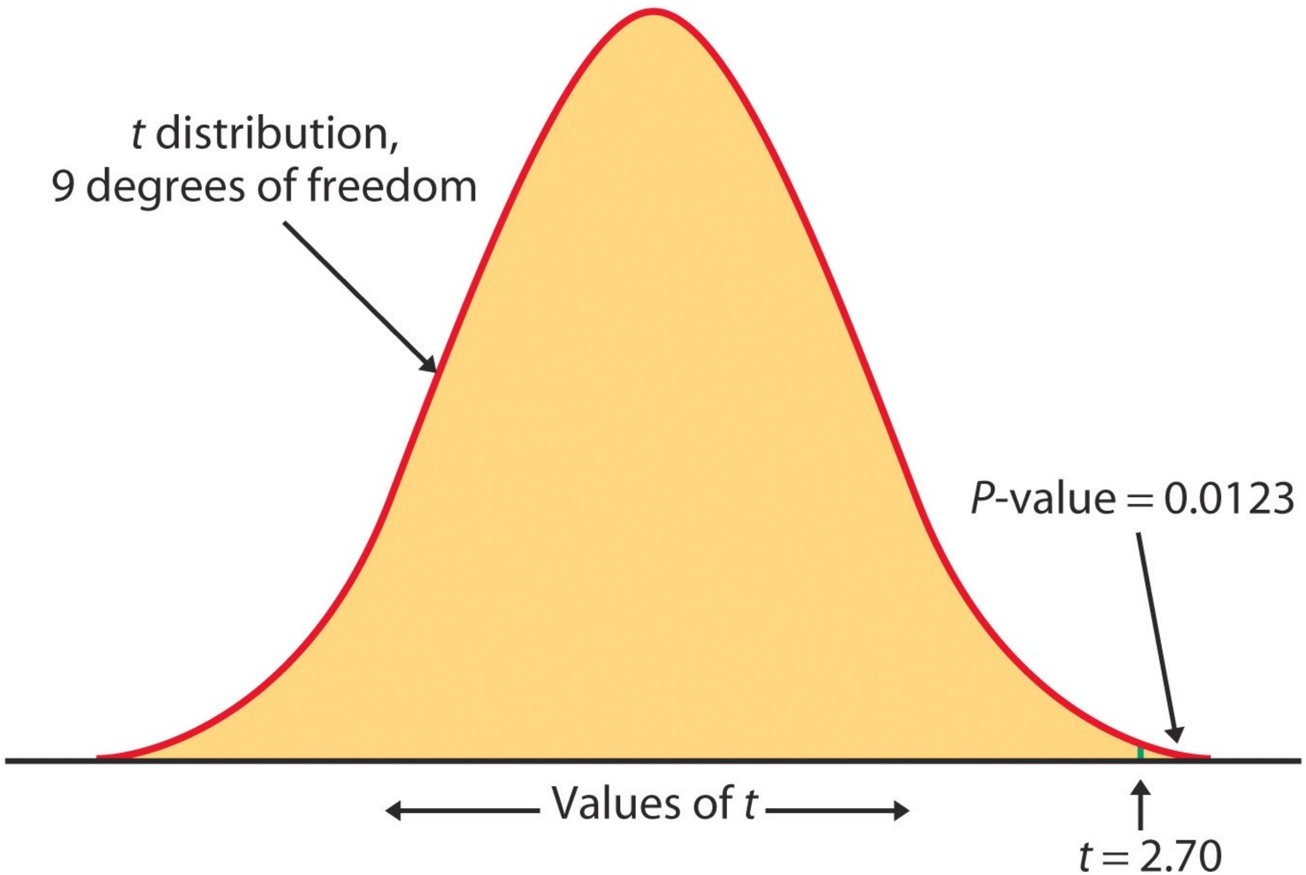
2. ***Test Statistic:*** $\bar{x} = 1.02, s = 1.196, t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = ?$
   (df = ?)

3. ***P-value:***
   P-value = $P(T \geq 2.70) = 0.0123$ *(using a computer)*
   P-value is between 0.01 and 0.02 since $t = 2.70$ is between $t^* = 2.398$ ($p = 0.02$) and $t^* = 2.821$ ($p = 0.01$) *(Table D)*

4. ***Conclusion:***
   Since the P-value is smaller than $\alpha = 0.05$, there is quite strong evidence that the new cola loses sweetness on average during storage at room temperature.

Matched Pair Case Study I

Sweetening Colas



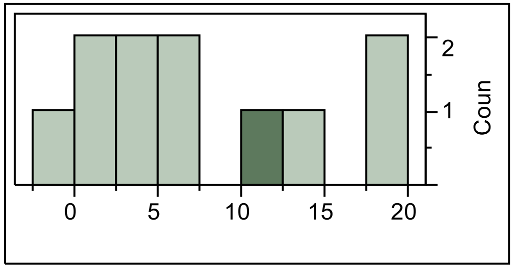
Matched pair case study II
Does lack of caffeine increase depression?

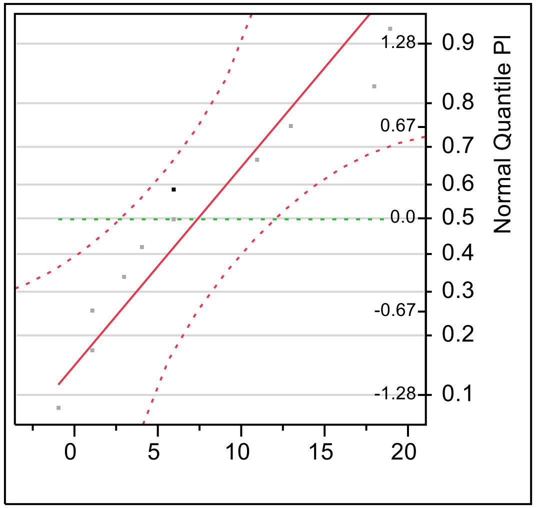
Subject	Depression with Caffeine	Depression with Placebo	Placebo - Cafeine
1	5	16	11
2	5	23	18
3	4	5	1
4	3	7	4
5	8	14	6
6	5	24	19
7	0	6	6
8	0	3	3
9	2	15	13
10	11	12	1
11	1	0	-1

Individuals diagnosed as caffeine-dependent are deprived of caffeine-rich foods and assigned to receive daily pills. Sometimes, the pills contain caffeine and other times they contain a placebo. Depression was assessed.

- There are 2 data points for each subject, but we'll only look at the difference.
- The sample distribution appears appropriate for a *t*-test.

11 "difference" data points.





Does lack of caffeine increase depression?

For each individual in the sample, we have calculated a difference in depression score (placebo minus caffeine).

There were 11 "difference" points, thus df = *?*

We calculate that $\bar{x}$= 7.36; s = 6.92

H_0: H_a:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = ?$$

P-value?

Conclusion:

JMP output for caffeine example

Treat it as 1 sample t-test using the difference: (placebo-caffeine),

Analyze: Distribution for the differences.

Test Mean

Hypothesized Value	0
Actual Estimate	7.36364
DF	10
Std Dev	6.9177

t Test

Test Statistic	3.5304
Prob > |t|	0.0054*
Prob > t	0.0027*
Prob < t	0.9973

Conducting a paired sample *t*-test on the raw data (caffeine and placebo)

Analyze: Matched Pairs

Matched Pairs
Difference: placebo-cafferine

placebo	11.3636	t-Ratio	3.530426
cafferine	4	DF	10
Mean Difference	7.36364	Prob > |t|	0.0054*
Std Error	2.08576	Prob > t	0.0027*
Upper 95%	12.011	Prob < t	0.9973
Lower 95%	2.71626		
N	11		
Correlation	0.4872		

Robustness of *t* procedures

- A confidence interval or significance test is called robust if the confidence level or *P*-value does not change very much when the conditions for use of the procedure are violated.

USING THE *t* PROCEDURES

- Except in the case of small samples, the condition that the data are an SRS from the population of interest is more important than the condition that the population distribution is Normal.
- *Sample size less than 15*: Use *t* procedures if the data appear close to Normal (roughly symmetric, single peak, and no outliers). If the data are clearly skewed or if outliers are present, do not use *t*.
- *Sample size at least 15*: The *t* procedures can be used, except in the presence of outliers or strong skewness.
- *Large samples*: The *t* procedures can be used, even for clearly skewed distributions when the sample is large, roughly $n \geq 40$.

What if the data are clearly not Normal?

- If lack of Normality is due to outliers, it may be legitimate to remove outliers if you have reason to think that they do not come from the same population as the other observations. ***But if an outlier appears to be "real data," you should not arbitrarily remove it.***
- In some settings, other standard distributions replace the Normal distributions as models for the overall pattern in the population. Statistical studies in these areas use families of right-skewed distributions rather than Normal distributions. There are inference procedures for the parameters of these distributions that replace the t procedures.
- Modern **bootstrap methods** and **permutation tests** use heavy computing to avoid requiring Normality or any other specific form of sampling distribution.
- Finally, there are other nonparametric methods, which do not assume any specific form for the distribution of the population. Unlike bootstrap and permutation methods, common nonparametric methods do not make use of the actual values of the observation.

Inference for non-normal distributions

What if the population is clearly non-normal and your sample is small?

- If the data are skewed, you can attempt to **transform** the variable to bring it closer to normality (e.g., logarithm transformation). The *t*- procedures applied to transformed data are quite accurate for even moderate sample sizes.

- A distribution other than a normal distribution might describe your data well. Many non-normal models have been developed to provide inference procedures too.

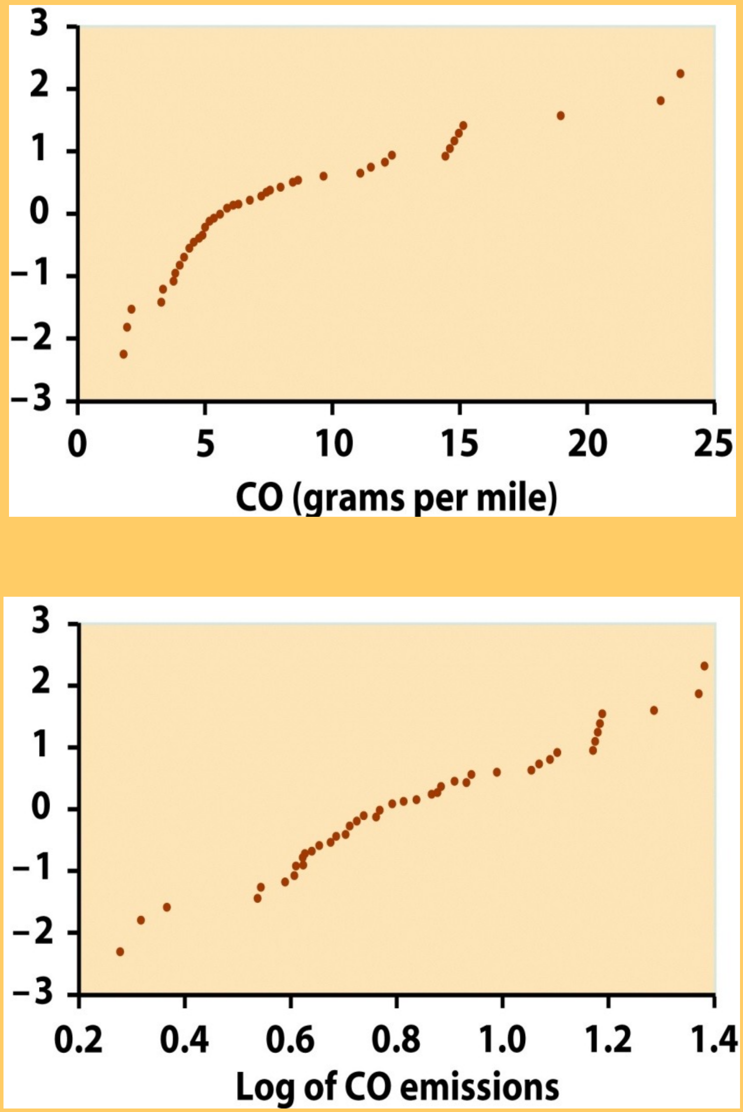
- You can always use a **distribution-free ("nonparametric")** inference procedure that does not assume any specific distribution for the population. But it is usually less powerful than distribution-driven tests (e.g., *t* test).

Transforming data

The most common transformation is the **logarithm (log),** which tends to pull in the right tail of a distribution.

Instead of analyzing the original variable *X*, we first compute the logarithms and analyze the values of log *X*.

However, we cannot simply use the confidence interval for the mean of the logs to deduce a confidence interval for the mean μ in the original scale.



Normal quantile plots for 46 car CO emissions