From Z to T

Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

Roadmap

Roadmap

Part 1 of the course looked at visualizing and describing data

- Continuous(histograms, box plots, mean, median, variance, standard deviation etc.)
- Categorical (bar charts, stacked bars, frequencies/percents marginal and conditional probabilities)

Part II introduced key concepts in probability and distributions

- Probability rules (independence, addition and decomposition, multiplication, Bayes theorem)
- Continuous distribution (Normal)
- Discreet distributions (Binomial and Poisson)
- Sampling variability, central limit theorem, CI and hypothesis testing

Part III will put these together and build your toolkit for statistical testing

Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

Roadmap

Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

In deciding what statistical test to use we will often be thinking about:

- The type of data (continuous vs categorical)
- How many groups we are comparing
- Is there an inherent relationship between the measurements (dependence)?
- Is there a theoretical distribution that is a good fit for the data?

Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

Reduced conditions for inference about a mean

Reduced conditions for inference about a mean

- Data is a SRS form a much larger population (really important)
- Observations follow a Normal distribution (some leeway)

Recap: Z testing

- We have been looking at variables that are continuous in nature
- For the last few lectures we have assumed that the population standard deviation (σ)was known to us
- We conducted the z-test and created CIs using this known σ
- Today we will generalize this framework to a more realistic setting where σ is unknown. We will use s, the sample standard deviation as an estimate of σ

Estimating the standard error based on the sample

- Previously, we knew the standard error of the mean to be

$$\frac{\sigma}{\sqrt{n}}$$

- Now, we don't know σ, so we estimate the standard error by

$$\frac{s}{\sqrt{n}}$$

where s is the sample standard deviation.

standard deviation vs standard error

- Variance σ^2: average squared deviation from the mean(absolute value)
- Standard Deviation (σ for a parameter or S for a sample) - square root of the variance of all observations: - on average how far do our values deviate from the samplemean
- Standard Error (SE): - The standard deviation of a statistic estimated from the data is the standard error of the statistic. - The standard error is $se = s/\sqrt{n}$ - The standard error is the sd of all sample means - Tells how close our test statistic is to the true value. - on average how far does our test statistic move from the true population mean?

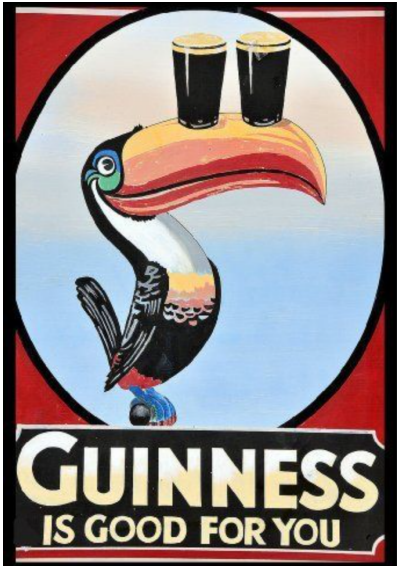
Recall the z-test!

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

the t-test and t-distribution

Meet the t-test!



T-test

Thanks to William Gosset (published anonymously) and the Guinness company's strategy of hiring statisticians, if we are interested in comparing mean values of a variable to a hypothesized null we can use a t-test.

Meet the t-test!

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

- What is the difference between z and t?
- The t-test is more variable than the z-test statistic because we have to substitute s for σ. Because s is a statistic, it varies across samples.
- Because of this substitution, the t-test will not follow a Normal(0, 1) distribution. It is *more* variable than the standard Normal. Thus, we need a distribution that is like the standard Normal but a little bit wider.

Variability and sample size

What do we know about our estimate of mean and variability as the sample size grows?

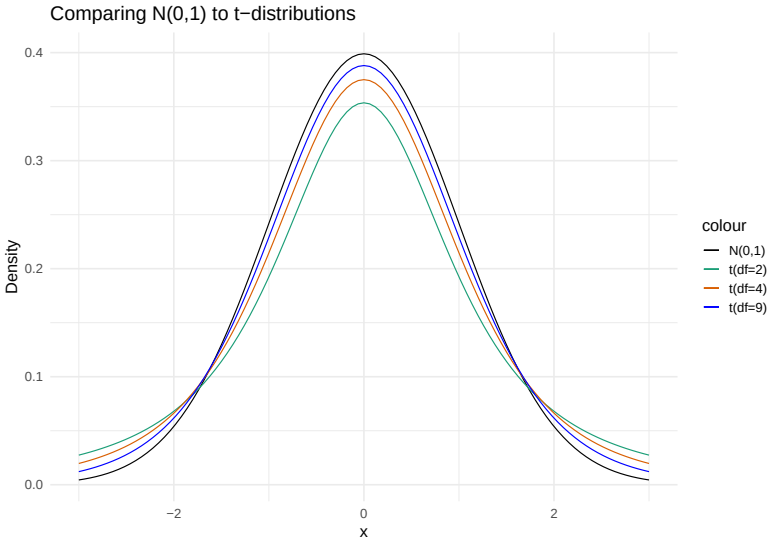
Introducing the t distribution

Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

- Like the standard Normal distribution, but wider.
- It's width depends on n, the sample size which determines the degrees of freedom

This is because as n increases, our estimate s gets better and better, and approaches σ. Thus, as n increases the t-distribution approaches a Normal(0, 1) distribution.

Introducing the t distribution



Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

Meet the t-test!

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

The one-sample t statistic has a t distribution with $n - 1$ degrees of freedom

You calculate the t statistic using $\bar{x}$ and s estimated from your sample, and n which is also a property of your sample, and μ_0 from the null hypothesis. Then compute the probability of observing a value of t or more extreme.

Assumptions of the t-test

- The dependent variable must be continuous (interval/ratio).
- The observations are independent of one another.
- Data come from a random sample of the underlying population.
- The dependent variable should be approximately normally distributed.
- The dependent variable should not contain any outliers.

Guess the R functions

Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

```
pt(q = , df = , lower.tail = )
qt(p = , df = , lower.tail = )
```

Which one would we use to calculate the p-value for a hypothesis test after we calculated the t-test statistic? `pt` or `qt`?

Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

Confidence intervals based on t

Calculating a confidence interval for the t-test

Draw an SRS of size n from a large population having unknown mean μ and unknown standard deviation σ. 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 $t*$.

Supposing we had $n = 100$, what is t^* for a 95% confidence interval?

```
qt(p = 0.975, df = 99)
```

Example: Testosterone and obesity in adolescent males (pg 422 B&M Ed 4)

Here are the data for $n = 25$ adolescent males between the ages of 14 and 20:

```
library(dplyr)
testosterone <- c(0.30, 0.24, 0.19, 0.17, 0.18, 0.23, 0.24, 0.06, 0.15,
                  0.17, 0.18, 0.17, 0.15, 0.12, 0.25, 0.25, 0.25, 0.32,
                  0.35, 0.37, 0.39, 0.46, 0.49, 0.42, 0.36)
dat_test <- data.frame(testosterone)
```

Example: Testosterone and obesity in adolescent males (pg 422 B&M Ed 4)

Use R to calculate a 95% confidence interval for testosterone. We can do this using `summarize`

```
dat_test %>% summarize(sample_mean = mean(testosterone),
                       sample_sd = sd(testosterone),
                       sample_size = length(testosterone),
                       sample_se = sample_sd/sqrt(sample_size))
```

```
##   sample_mean sample_sd sample_size  sample_se
## 1      0.2584 0.1115303          25 0.02230605
```

Example: Testosterone and obesity in adolescent males (pg 422 B&M Ed 4)

Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

We still need the t^* value:

```
t_star <- qt(p = 0.975, df = 24)
t_star
```

```
## [1] 2.063899
```

Example: Testosterone and obesity in adolescent males (pg 422 B&M Ed 4)

Expand the previous code chunk to calculate the margin of error (which uses the critical t^* value), and then calculate the lower and upper CI

```
dat_test %>% summarize(sample_mean = mean(testosterone),
                       sample_sd = sd(testosterone),
                       sample_size = length(testosterone),
                       sample_se = sample_sd/sqrt(sample_size),
                       margin_of_error = sample_se*t_star,
                       lower_CI = sample_mean - margin_of_error,
                       upper_CI = sample_mean + margin_of_error)
```

```
##   sample_mean sample_sd sample_size  sample_se margin_of_error  lower_CI
## 1      0.2584 0.1115303          25 0.02230605      0.04603743 0.2123626
##    upper_CI
## 1 0.3044374
```

Hypothesis testing with unknown σ using the t-test

Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

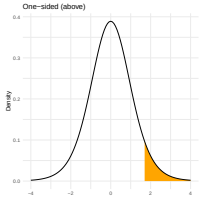
Draw an SRS of size n from a large population having unknown mean μ and unknown standard deviation σ. To test the hypothesis $H_0 : \mu = \mu_0$, calculate the t statistic:

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

Hypothesis testing with unknown σ using the t-test

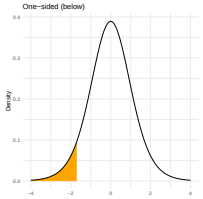
In terms of a variable T having the $t(n-1)$ distribution, the p-value for a test of H_0 against

H_a: $\mu > \mu_0$ is $P(T \geq t)$



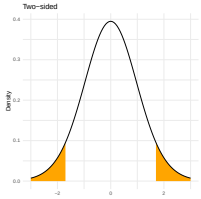
Hypothesis testing with unknown σ using the t-test

H_a: $\mu < \mu_0$ is $P(T \leq t)$



Hypothesis testing with unknown σ using the t-test

H_a: $\mu \neq \mu_0$ is $2 \times P(T \geq |t|)$



Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

Example and t-testing in R

Example of a t-test (pg 426 B&M Ed 4)

Here are 18 measures of pulse wave velocity (PWV) from a sample of children diagnosed with progeria, a genetic disorder that produces rapid aging.

```
pwv <- c(18.8, 17.6, 17.5, 16.0, 14.8, 14.1, 13.7, 13.1, 12.9,
         12.9, 12.4, 10.1, 9.3,  9.1,  8.3,  8.3,  7.9,  7.2)

pwv_dat <- data.frame(pwv)
```

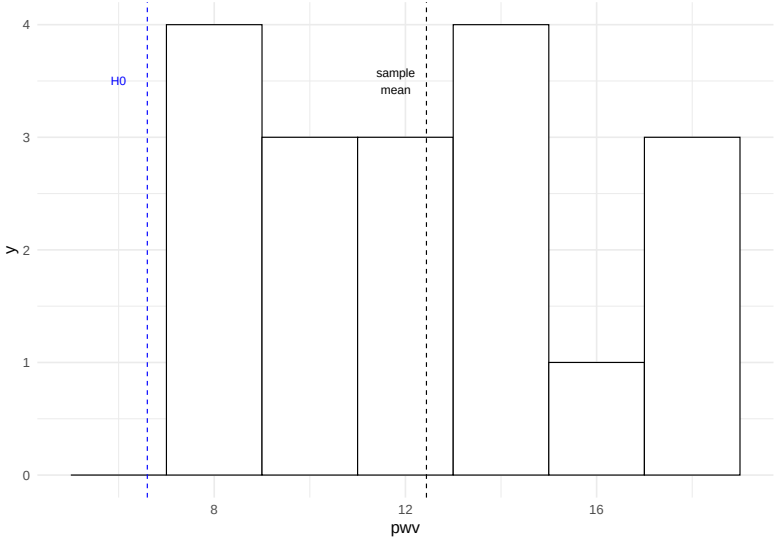
Example of a t-test (pg 426 B&M Ed 4)

Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

pwv measures greater than 6.6 are considered abnormally high. We would like to test the hypothesis that the mean for this subset of children is abnormally high.

That is: $H_0 : \mu = 6.6$ and $H_a : \mu > 6.6$

Look at the data and see if there is evidence against the null hypothesis



Example of a one-sided t-test (pg 426 B&M Ed 4)


```
pwv_dat %>%
  summarize(sample_mean = mean(pwv),
            sample_sd = sd(pwv),
            sample_size = length(pwv),
            sample_se = sample_sd/sqrt(sample_size),
            t_test = (sample_mean - 6.6)/sample_se,
            p_value = 1 - pt(t_test, df = sample_size - 1))
```

```
##   sample_mean sample_sd sample_size sample_se   t_test      p_value
## 1    12.44444  3.637747          18 0.8574252 6.816273 1.501248e-06
```

There's a function for that...

Rather than doing the test using summarize, we could have R do it for us using `t.test`:

```
t.test(x = pwv_dat %>% pull(pwv), alternative = "greater", mu = 6.6)
```

```
##
##  One Sample t-test
##
## data:  pwv_dat %>% pull(pwv)
## t = 6.8163, df = 17, p-value = 1.501e-06
## alternative hypothesis: true mean is greater than 6.6
## 95 percent confidence interval:
##  10.95286      Inf
## sample estimates:
## mean of x
##  12.44444
```

Robustness of t procedures

- A confidence interval or hypothesis 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.
- In particular, how robust are the procedures against non-Normality?
- The t procedures are quite robust against non-Normality of the population except when outliers or strong skewness are present.
- The t procedures are not robust against outliers unless the sample size is sufficiently large.

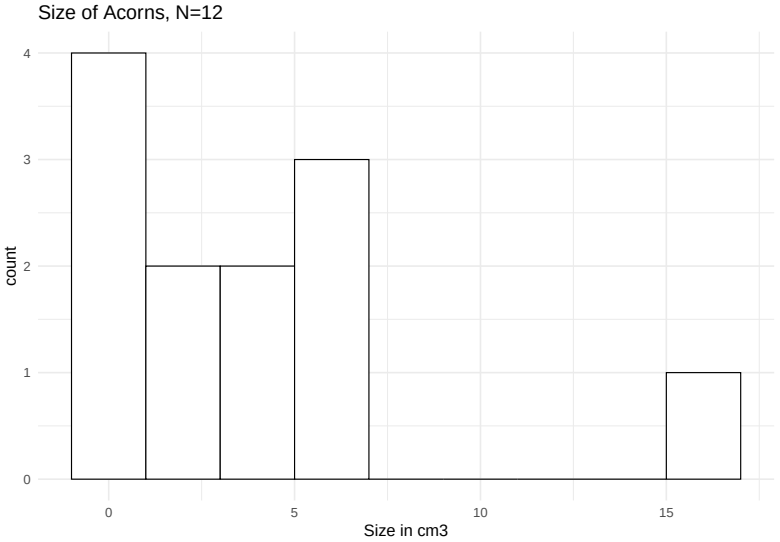
Checking assumptions

- Always plot your data first:
  - Are there any outliers
  - Is the distribution of the data skewed?

Guidelines for using the t procedures

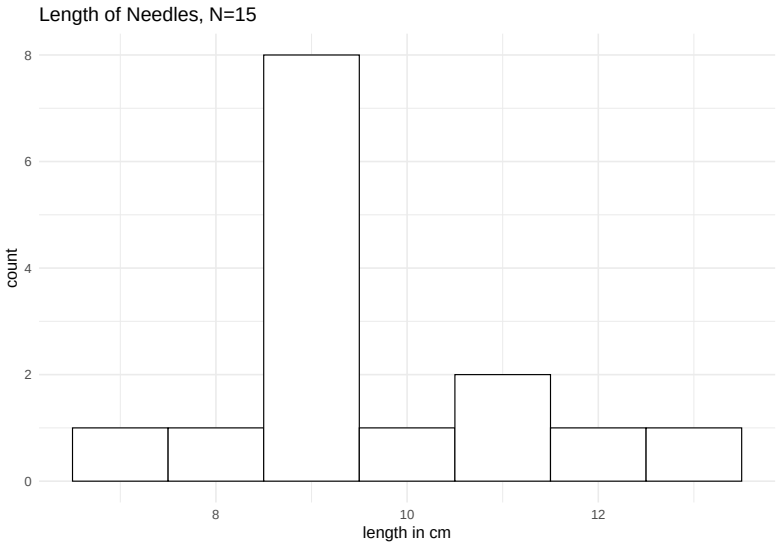
- The SRS condition is more important that the Normality condition
- If n $<$ 15: Use t procedures if the data appear close to Normal (at least roughly symmetric, single peak, no outliers). If the data are skewed or there are outliers, don't use t.
- Moderate sample size $>$ 15: The t procedures can be used except in the presence of outliers or strong skewness
- Large sample size, roughly $n \geq 40$: The t procedures can be used even for strongly skewed distributions when the sample is large, roughly $n \geq 40$

Example : Can we use t?



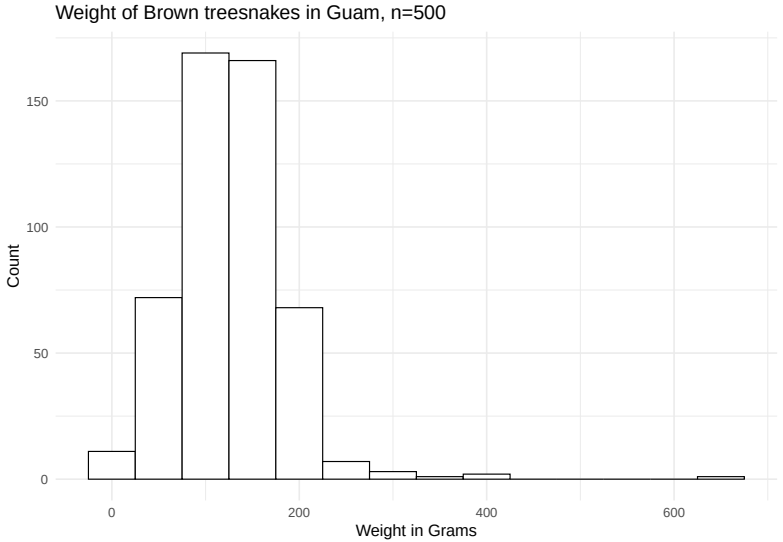
Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

Example : Can we use t?



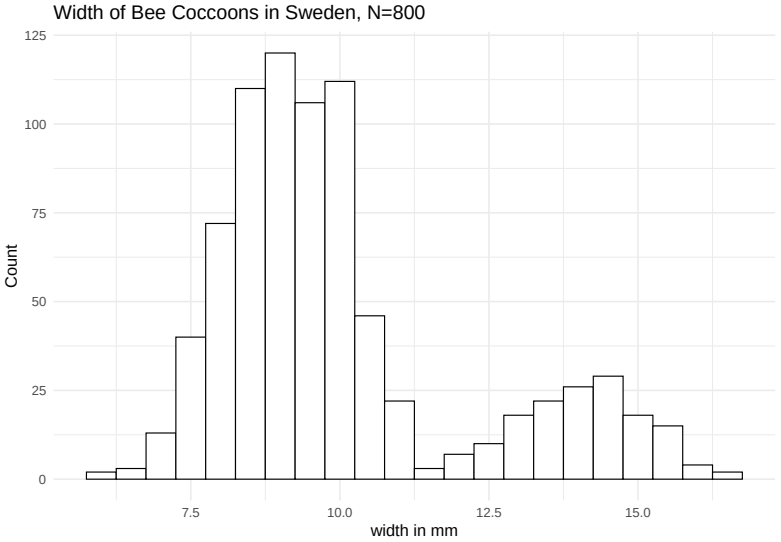
Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

Example : Can we use t?



Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

Example: Can we use t?



Roadmap
Reduced conditions for inference about a mean
the t-test and t-distribution
Confidence intervals based on t
Example and t-testing in R

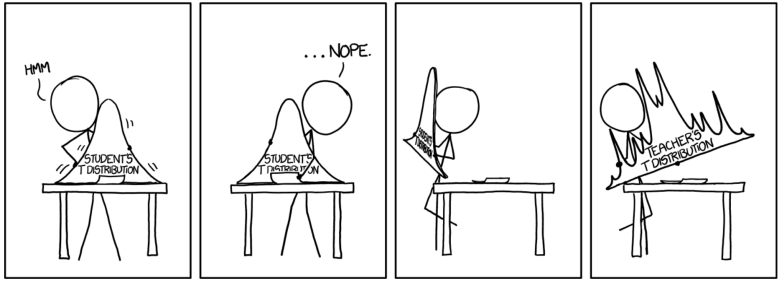
Recap

- We use a z test when the population sd or σ is known
- We use a t test when the population sd is being estimated by s
- We compare a z test statistic to a N(0,1) distribution to calculate a probability (p-value)
- We compare the t test statistic to a t distribution with n-1 degrees of freedom
- When n is large, the t distribution is very close to the N(0,1) distribution.

New Code

pt(q = , df = , lower.tail =)

qt(p = , df = , lower.tail =)

t.test(x =, alternative = , mu =)

Parting Humor