## Friday the 13th

1990 - 1992 UK traffic flow on Friday 13th and previous Friday, Friday 6th. Assume traffic flow on given day at locations 1 and 2 is independent.

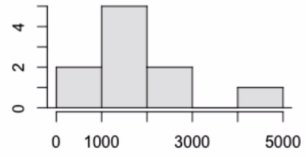
| | type | date | 6th | 13th | diff | location |
|---|---|---|---|---|---|---|
| 1 | traffic | 1990, July | 139246 | 138548 | 698 | loc1 |
| 2 | traffic | 1990, July | 134012 | 132908 | 1104 | loc2 |
| 3 | traffic | 1991, September | 137055 | 136018 | 1037 | loc1 |
| 4 | traffic | 1991, September | 133732 | 131843 | 1889 | loc2 |
| 5 | traffic | 1991, December | 123552 | 121641 | 1911 | loc1 |
| 6 | traffic | 1991, December | 121139 | 118723 | 2416 | loc2 |
| 7 | traffic | 1992, March | 128293 | 125532 | 2761 | loc1 |
| 8 | traffic | 1992, March | 124631 | 120249 | 4382 | loc2 |
| 9 | traffic | 1992, November | 124609 | 122770 | 1839 | loc1 |
| 10 | traffic | 1992, November | 117584 | 117263 | 321 | loc2 |

## Friday the 13th hypothesis test

1. Set hypotheses
   $H_0$: Average traffic flow on Friday 6th and 13th are equal.
   $\mu_{diff} = 0$
   $H_A$: Average traffic flow on Friday 6th and 13th are different.
   $\mu_{diff} \neq 0$
2. Calculate point estimate
   $\bar{x}_{diff} = 1836$

## Friday the 13th hypothesis test

3. Check conditions
   a. Independence: assumed
   b. Sample size/skew




We do not know $\sigma$ and n is too small to assume s is a reliable estimate for $\sigma$

## *s* as an estimate for $\sigma$

- Sample std. dev. *s* is a biased estimate of $\sigma$
- Specifically: *s* is on average smaller than $\sigma$

Bias of *s* is small if the sample size is large (>30 or so), so we were able to ignore this in early examples

Small sample sizes require a correction, however

## The t distribution

Replaces standard normal distribution in calculations

Handles unknown population standard error

Has a single parameter: degrees of freedom (df)

1 less than sample size


## The t distribution test statistic

The test statistic for inference on a small sample (n < 50) mean is the $T$ statistic with df = n - 1:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

$$SE = \frac{\text{sample standard deviation}}{\sqrt{\text{sample size}}}$$

## Friday the 13th hypothesis test

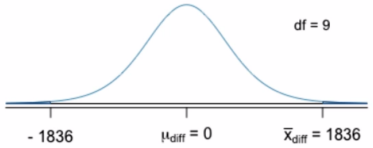
5. Draw a conclusion

Since p = 0.0008, the data provides evidence of a difference between traffic flow on Friday the 6th and Friday the 13th

## Friday the 13th hypothesis test

4. Calculate the test statistic and p-value

$$SE = \frac{\bar{s}_{diff}=1176}{\sqrt{10}} = 372 \quad T_{df} = \frac{\bar{x}_{diff}=1836 - 0}{372} = 4.94$$




```
> 2 * pt(4.94, df = 10 - 1, lower.tail = FALSE)
[1] 0.0008022394
```

N = 12 $\bar{x}$ = 18.64 SD = 7.19

## t distribution confidence intervals

Similar form to normal distribution:

point estimate ± t* × standard error

95% confidence

For Friday the 6th vs. 13th:

```
> qt(.025, df = 10 - 1, lower.tail = FALSE)
[1] 2.262157
```

wider than normal dis

1836 ± 2.262157 × 372

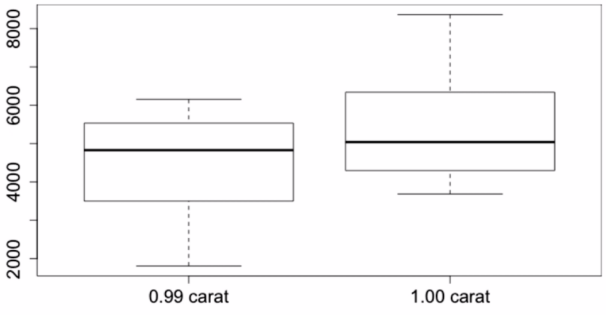
(994, 2678) True diff between 2 days

## When to use the t distribution

- Whenever we are investigating the mean of a numerical variable and the population SD is known (don't know)
- Sample size cutoffs (e.g. 30, 50) -- below these cutoffs, t distribution is definitely needed
- Always more accurate than normal distribution for unknown population SD, but at large sample size the difference is negligible

# Difference of 2 means

## Price per carat of diamonds




Is a 0.99 carat diamond a better deal?

## Parameter and point estimate

Parameter of interest

Difference between the average price per carat of **all** 0.99 carat and 1 carat diamonds

$\mu_{0.99\text{-carat}} - \mu_{1.00\text{-carat}}$

Point estimate

Difference between the average price per carat of **sampled** 0.99 carat + 1 carat diamonds

$\bar{x}_{0.99\text{-carat}} - \bar{x}_{1.00\text{-carat}}$

## Hypotheses

Is the average price per carat of 1 carat diamonds higher than the average price per carat of 0.99 carat diamonds?

$H_0$: $\mu_{0.99\text{-carat}} - \mu_{1.00\text{-carat}} = 0$

$H_A$: $\mu_{0.99\text{-carat}} - \mu_{1.00\text{-carat}} < 0$

## Test statistic

For inference on the difference of two means where $\sigma_1$ and $\sigma_2$ are unknown, use the t statistic

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{\text{standard error}}$$

But how do we calculate standard error when there are two standard deviations: $\sigma_1$ and $\sigma_2$?

And how do we calculate the degrees of freedom with two sample sizes: $n_1$ and $n_2$?

## T statistic with $\sigma_1$ and $\sigma_2$

Standard error: $$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Degrees of freedom: $$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

## T statistic for diamonds

Sample statistics:

| | 0.99 carat | 1.00 carat |
|---|---|---|
| $\bar{x}$ | 4451 | 5343 |
| s | 1332 | 1222 |
| n | 23 | 30 |

T statistic and p value:

```
> se <- sqrt(s1^2/n1 + s2^2/n2)
> t <- ((m1 - m2) - 0) / se
> t      mean
[1] -2.503838
> df <- (s1^2/n1 + s2^2/n2)^2 /
+         ((s1^2/n1)^2/(n1-1) +
+          (s2^2/n2)^2/(n2-1))
> df
[1] 45.25681
> pt(t, df, lower.tail = TRUE)
[1] 0.007978864
```

## Confidence Interval

Same form as usual:

point estimate ± t* × standard error

For the diamond example:

```
> critical.value <- qt(0.975, df)
> margin <- critical.value * se
> (m1 - m2) + c(-margin, +margin)
[1] -1609.4183  -174.5817
```

We are 95% confident that the price per carat of a .99 carat diamond is between $1609 and $174 less than that of a 1 carat diamond

## Conditions

- Similar conditions apply as for one-sample tests
  - Independent random sampling
  - For small samples, distributions not badly skewed
- Two-sample t-tests are most reliable when
  - Both distributions are not badly skewed
  - The two distributions have similar standard deviations
  - The two samples are similar in size