

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2006, The Johns Hopkins University and Brian Caffo. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

Outline

1. Define the Chi-squared and t distributions
2. Derive confidence intervals for the variance
3. Illustrate the likelihood for the variance
4. Derive t confidence intervals for the mean
5. Derive the likelihood for the effect size

Confidence intervals

- Previously, we discussed creating a confidence interval using the CLT
- Now we discuss the creation of better confidence intervals for small samples using Gosset's t distribution
- To discuss the t distribution we must discuss the Chi-squared distribution
- Throughout we use the following general procedure for creating CIs
 - a. Create a **Pivot** or statistic that does not depend on the parameter of interest
 - b. Solve the probability that the pivot lies between bounds for the parameter

The Chi-squared distribution

- Suppose that S^2 is the sample variance from a collection of iid $N(\mu, \sigma^2)$ data; then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

which reads: follows a Chi-squared distribution with $n-1$ degrees of freedom

- The Chi-squared distribution is skewed and has support on 0 to ∞
- The mean of the Chi-squared is its degrees of freedom
- The variance of the Chi-squared distribution is twice the degrees of freedom

Confidence interval for the variance

Note that if $\chi_{n-1,\alpha}^2$ is the α quantile of the Chi-squared distribution then

$$\begin{aligned} 1 - \alpha &= P \left(\chi_{n-1,\alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{n-1,1-\alpha/2}^2 \right) \\ &= P \left(\frac{(n-1)S^2}{\chi_{n-1,1-\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{n-1,\alpha/2}^2} \right) \end{aligned}$$

So that

$$\left[\frac{(n-1)S^2}{\chi_{n-1,1-\alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1,\alpha/2}^2} \right]$$

is a $100(1 - \alpha)\%$ confidence interval for σ^2

Notes about this interval

- This interval relies heavily on the assumed normality
- Square-rooting the endpoints yields a CI for σ
- It turns out that

$$(n - 1)S^2 \sim \text{Gamma}\{(n - 1)/2, 2\sigma^2\}$$

which reads: follows a gamma distribution with shape $(n - 1)/2$ and scale $2\sigma^2$

- Therefore, this can be used to plot a likelihood function for σ^2

Example

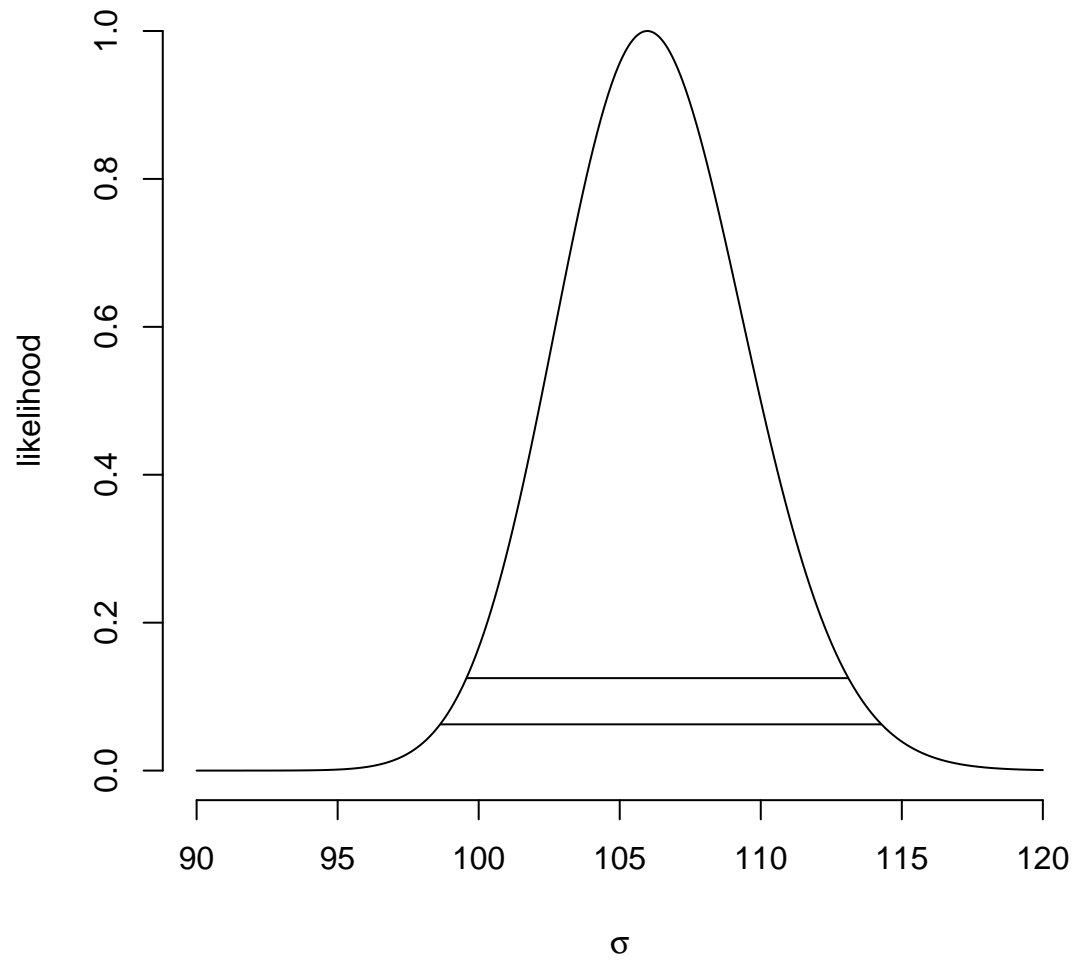
A recent study of 513 organo-lead manufacturing workers reported an average total brain volume of $1,150.315 \text{ cm}^3$ with a standard deviation of 105.977. Assuming normality of the underlying measurements, calculate a confidence interval for the population variation in total brain volume.

Example continued

```
##CI for the variance
s2 <- 105.977 ^ 2
n <- 513
alpha <- .05
qtiles <- qchisq(c(alpha/2, 1 - alpha/2),
                 n - 1)
ival <- rev((n - 1) * s2 / qtiles)
##interval for the sd
sqrt(ival)
[1] 99.86484 112.89216
```

Plot the likelihood

```
sigmaVals <- seq(90, 120, length = 1000)
likeVals <- dgamma((n - 1) * s2,
                   shape = (n - 1)/2,
                   scale = 2*sigmaVals^2)
likeVals <- likeVals / max(likeVals)
plot(sigmaVals, likeVals)
lines(range(sigmaVals[likeVals >= 1 / 8]),
      c(1 / 8, 1 / 8))
lines(range(sigmaVals[likeVals >= 1 / 16]),
      c(1 / 16, 1 / 16))
```



Gosset's t distribution

- Invented by William Gosset (under the pseudonym “Student”) in 1908
- Has thicker tails than the normal
- Is indexed by a degrees of freedom; gets more like a standard normal as df gets larger
- Is obtained as

$$\frac{Z}{\sqrt{\frac{\chi^2}{df}}}$$

where Z and χ^2 are independent standard normals and Chi-squared distributions respectively

Result

- Suppose that (X_1, \dots, X_n) are iid $N(\mu, \sigma^2)$, then:
 - a. $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is standard normal
 - b. $\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} = S/\sigma$ is the square root of a Chi-squared divided by its df

- Therefore

$$\frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{S/\sigma} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows Gosset's t distribution with $n - 1$ degrees of freedom

Confidence intervals for the mean

- Notice that the t statistic is a pivot, therefore we use it to create a confidence interval for μ
- Let $t_{df,\alpha}$ be the α^{th} quantile of the t distribution with df degrees of freedom

$$1 - \alpha = P \left(-t_{n-1,1-\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1,1-\alpha/2} \right)$$

$$= P \left(\bar{X} - t_{n-1,1-\alpha/2}S/\sqrt{n} \leq \mu \leq \bar{X} + t_{n-1,1-\alpha/2}S/\sqrt{n} \right)$$

- Interval is $\bar{X} \pm t_{n-1,1-\alpha/2}S/\sqrt{n}$

Note's about the t interval

- The t interval technically assumes that the data are iid normal, though it is robust to this assumption
- It works well whenever the distribution of the data is roughly symmetric and mound shaped
- Paired observations are often analyzed using the t interval by taking differences
- For large degrees of freedom, t quantiles become the same as standard normal quantiles; therefore this interval converges to the same interval as the CLT yielded

- For skewed distributions, the spirit of the t interval assumptions are violated
- Also, for skewed distributions, it doesn't make a lot of sense to center the interval at the mean
- In this case, consider taking logs or using a different summary like the median
- For highly discrete data, like binary, other intervals are available

Sleep data

In R typing `data(sleep)` brings up the sleep data originally analyzed in Gosset's Biometrika paper, which shows the increase in hours for 10 patients on two soporific drugs. R treats the data as two groups rather than paired.

Patient	g1	g2	diff
1	0.7	1.9	1.2
2	-1.6	0.8	2.4
3	-0.2	1.1	1.3
4	-1.2	0.1	1.3
5	-0.1	-0.1	0.0
6	3.4	4.4	1.0
7	3.7	5.5	1.8
8	0.8	1.6	0.8
9	0.0	4.6	4.6
10	2.0	3.4	1.4

```
data(sleep)
g1 <- sleep$extra[1 : 10]
g2 <- sleep$extra[11 : 20]
difference <- g2 - g1
mn <- mean(difference)#1.67
s <- sd(difference)#1.13
n <- 10
mn + c(-1, 1) * qt(.975, n-1) * s / sqrt(n)
t.test(difference)$conf.int
[1] 0.7001142 2.4598858
```

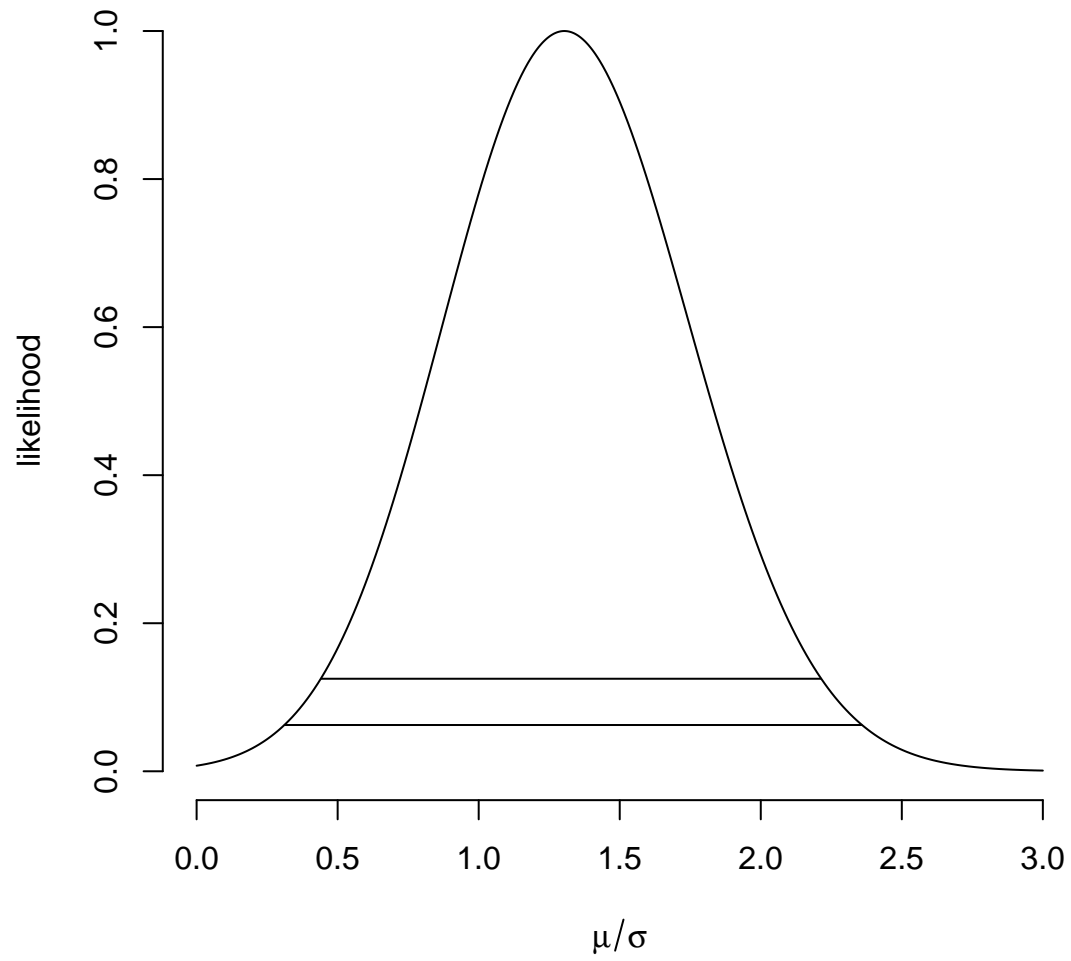
The non-central t distribution

- If X is $N(\mu, \sigma^2)$ and χ^2 is a Chi-squared random variable with df degrees of freedom then $\frac{X/\sigma}{\sqrt{\frac{\chi^2}{df}}}$ is called a **non-central t** random variable with non-centrality parameter μ/σ
- Note that
 - a. \bar{X} is $N(\mu, \sigma^2/n)$
 - b. $(n - 1)S^2/\sigma^2$ is Chi-squared with $n - 1$ df
- Then $\sqrt{n}\bar{X}/S$ is non-central t with non-centrality parameter $\sqrt{n}\mu/\sigma$
- We can use this to create a likelihood for μ/σ , the **effect size**

Some code

Starting after the code for the t interval

```
tStat <- sqrt(n) * mn / s
esVals <- seq(0, 1, length = 1000)
likVals <- dt(tStat, n - 1, ncp = sqrt(n) * esVals)
likVals <- likVals / max(likVals)
plot(esVals, likVals, type = "l")
lines(range(esVals[likVals>1/8]), c(1/8,1/8))
lines(range(esVals[likVals>1/16]), c(1/16,1/16))
```



The profile likelihood

- To obtain a likelihood for μ alone, the preferred method is called **profiling**
- The profile likelihood gets its name because the result is like the shadow you would get if you were to shine a light on the two-dimensional likelihood for μ and σ
- The profile likelihood for parameter value μ_0 is obtained by maximizing the joint likelihood for σ with μ fixed at μ_0
- This process is repeated for lots of values of μ_0

Calculating the profile likelihood

- The joint likelihood with μ fixed at μ_0 is

$$\begin{aligned} &\propto \prod_{i=1}^n \sigma^{-1/2} \exp \left\{ -(x_i - \mu_0)^2 / 2\sigma^2 \right\} \\ &= \sigma^{-n/2} \exp \left\{ - \sum_{i=1}^n (x_i - \mu_0)^2 / 2\sigma^2 \right\} \end{aligned}$$

- With μ_0 fixed, the maximum likelihood estimator for σ^2 is $\sum_{i=1}^n (x_i - \mu_0)^2 / n$ (homework)
- Plugging this back into the likelihood we get

$$\left(\sum_{i=1}^n (x_i - \mu_0)^2 / n \right)^{-n/2} \exp(-n/2)$$

Continued

- Therefore, removing multiplicative constants, the profile likelihood is

$$\left(\sum_{i=1}^n (x_i - \mu)^2 \right)^{-n/2}$$

- Note that this is clearly maximized at $\mu = \bar{X}$, the same as the ML estimate for μ for the complete likelihood

Some code

```
muVals <- seq(0, 3, length = 1000)
likVals <- sapply(muVals,
                  function(mu){
                    (sum((difference - mu)^2) /
                     sum((difference - mn)^2)) ^ (-n/2)
                  })
plot(muVals, likVals, type = "l")
lines(range(muVals[likVals>1/8]), c(1/8,1/8))
lines(range(muVals[likVals>1/16]), c(1/16,1/16))
```

