Lecture 9

Ciprian Crainiceanu

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

September 8, 2020

Table of contents

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**: a function of data and parameters whose distribution does not depend on the parameter of interest
  b. Calculate the probability that the pivot lies in a particular interval
  c. Re-express the confidence interval in terms of (random) bounds on the parameter of interest

The Chi-squared distribution

- If $X_1, \ldots, X_n$ are independent $N(0,1)$ rvs then

$$V_n = \sum_{i=1}^{n} X_i^2$$

  has a Chi-squared distribution with n degrees of freedom
- We denote $V_n \sim \chi^2_n$
- The Chi-squared distribution is skewed and has support $(0, \infty)$
- The mean of the Chi-squared is its degrees of freedom
- The variance of the Chi-squared distribution is twice the degrees of freedom

Chi-squared distribution

- If $X \sim N(0,1)$ then

$$V = X^2 \sim \chi_1^2$$

Denote by $\Phi(x) = P(X \leq x)$ the cdf of the Normal distribution

$$\begin{aligned} F_V(v) &= P(X^2 \leq v) \\ &= P(-\sqrt{v} \leq X \leq \sqrt{v}) \\ &= \Phi(\sqrt{v}) - \Phi(-\sqrt{v}) \\ &= 2\Phi(\sqrt{v}) - 1 \end{aligned}$$

Chi-squared distribution

Recall that the pdf of the $N(0,1)$ is $\phi(x) = \Phi'(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$
Then the pdf of the χ_1^2 distribution is

$$
\begin{aligned}
f_V(v) &= F_V'(v) = 2\frac{1}{2\sqrt{v}}\Phi'(\sqrt{v}) \\
&= \frac{1}{\sqrt{2\pi v}} e^{-v/2}
\end{aligned}
$$

- The χ_1^2 distribution is the $\mathrm{Gamma}(1/2, 1/2)$ distribution
- $E(V) = 1$, $\mathrm{Var}(V) = 2$
- $E(V_n) = \sum_{i=1}^n E(X_i^2) = n$
- $\mathrm{Var}(V_n) = \sum_{i=1}^n \mathrm{Var}(X_i^2) = 2n$
- It can be shown that $V_n \sim \mathrm{Gamma}(n/2, 1/2) = \chi_n^2$

The Chi-squared distribution

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

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

Sketch of proof: $(X_i - \mu)/\sigma \sim N(0,1)$ and are independent

- $\sum_{i=1}^n \frac{(X_i-\mu)^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i-\bar{X}_n)^2}{\sigma^2} + \frac{n(\bar{X}_n-\mu)^2}{\sigma^2}$
- $\sum_{i=1}^n \frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2_n$, $\frac{n(\bar{X}_n-\mu)^2}{\sigma^2} \sim \chi^2_1$
- It will be shown that $\sum_{i=1}^n \frac{(X_i-\bar{X}_n)^2}{\sigma^2} \amalg \frac{n(\bar{X}_n-\mu)^2}{\sigma^2}$
- The only distribution of $\sum_{i=1}^n \frac{(X_i-\bar{X}_n)^2}{\sigma^2}$ that satisfies this is a χ^2_{n-1} (using a characteristic function argument)

Independence of the Normal mean and deviations from the mean

Let $X_1, \ldots, X_n \sim N(0,1)$ independent: then the sample mean $\bar{X}_n$ is independent of the vector of deviations from the mean $(X_1 - \bar{X}_n, \ldots, X_n - \bar{X}_n)$

Sketch of proof: $(X_1, \ldots, X_n)$ is a multivariate normal vector

- $(\bar{X}_n, X_1 - \bar{X}_n, \ldots, X_n - \bar{X}_n)$ is a multivariate normal random vector because it is a linear transformation of the vector $(X_1, \ldots, X_n)$
- It is enough to show $\mathrm{Cov}(\bar{X}_n, X_1 - \bar{X}_n) = 0$
- Implies $\bar{X}_n$ is independent of any function of $(X_1 - \bar{X}_n, \ldots, X_n - \bar{X}_n)$, including S^2

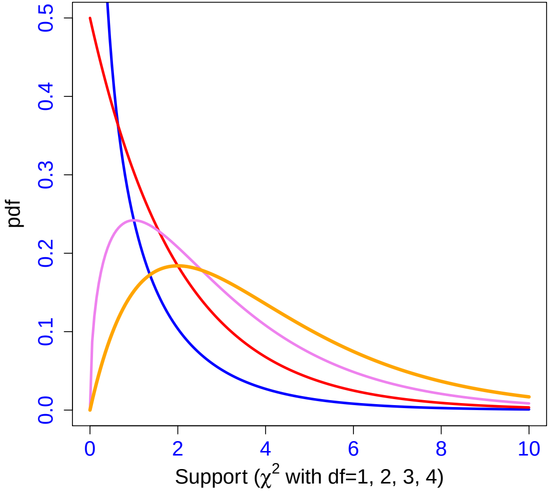
Covariance of the mean with the deviations from the mean

$$
\begin{aligned}
\mathrm{Cov}(\bar{X}_n, X_1 - \bar{X}_n) &= E\{\bar{X}_n(X_1 - \bar{X}_n)\} - E(X_1)E(X_1 - \bar{X}_n) \\
&= E(\bar{X}_n X_1) - E(\bar{X}_n^2) \\
&= E(\bar{X}_n X_1) - \{\mathrm{Var}(\bar{X}_n) + E^2(\bar{X}_n)\} \\
&= E(\bar{X}_n X_1) - (\sigma^2/n + \mu^2)
\end{aligned}
$$

We just need to show that $E(\bar{X}_n X_1) = \sigma^2/n + \mu^2$

$$
\begin{aligned}
E(\bar{X}_n X_1) &= \frac{1}{n}\sum_{i=1}^{n} E(X_1 X_i) \\
&= \frac{1}{n}\{E(X_1^2) + \sum E(X_1)E(X_i)\} \\
&= \frac{1}{n}\{\mathrm{Var}(X_1) + E^2(X_1) + \sum E(X_1)E(X_i)\} \\
&= \sigma^2/n + \mu^2
\end{aligned}
$$

Chi-squared distributions



R: Chi-squared quantiles

```
##quantiles of a chi-square distribution
n=4
alpha <- .05
qchisq(c(alpha/2, 1 - alpha/2),n)

##results
[1]  0.484 11.143
```

- For large n: the approximation $\chi^2_n \approx N(n, 2n)$ works very well for estimating the quantiles
- For large n: $(n-1)S_n/\sigma^2 \approx N(n-1, 2n-2)$ irrespective to the distribution of X (CLT)

Confidence interval for the variance

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

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

So that

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

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

Example

- A recent study 513 of organo-lead manufacturing workers reported an average total brain volume of $1,150.315\mathrm{cm}^3$ 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
```

Notes about this interval

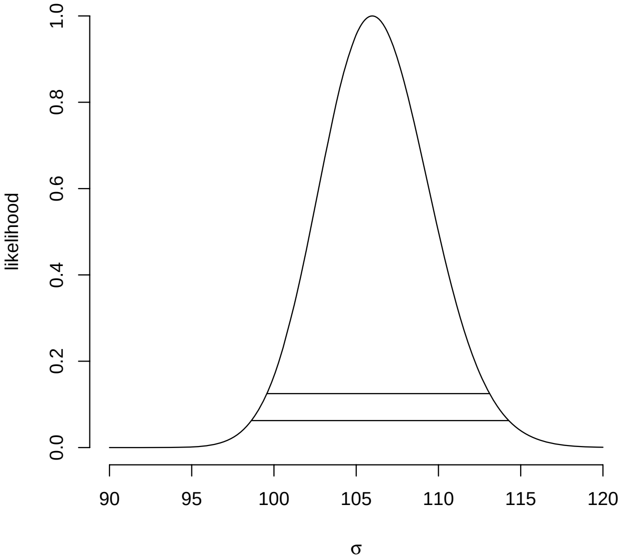
- This interval relies heavily on the assumed normality
- Square-rooting the endpoints yields a CI for σ
- It turns out that

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

  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

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, type = "l")
lines(range(sigmaVals[likeVals >= 1 / 8]),
      c(1 / 8, 1 / 8))
lines(range(sigmaVals[likeVals >= 1 / 16]),
      c(1 / 16, 1 / 16))
```


likelihood
0.0
0.2
0.4
0.6
0.8
1.0
90
95
100
105
110
115
120
σ

Proof of the variance likelihood result

If $X/a \sim \mathrm{Gamma}(\alpha, \beta)$ then $X \sim \mathrm{Gamma}(\alpha, a\beta)$
Let $F_X(x)$ be the cdf of X. Then
$F_{X/a}(x) = P(X \leq ax) = F(ax)$
Then the pdf

$$
\begin{aligned}
F'_{X/a}(x) &= aF'(ax) \\
&= a\frac{\beta^\alpha}{\Gamma(\alpha)}(ax)^{\alpha-1}e^{-a\beta x} \\
&= \frac{(a\beta)^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-(a\beta)x}
\end{aligned}
$$

Student's t distribution

- Invented by William Gosset (under the pseudonym "Student") in 1908
- Has thicker tails than the normal
- Is indexed by 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 $(X_1, \ldots, X_n)$ 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
  c. $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ and S/σ are independent (why?)
- Therefore

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

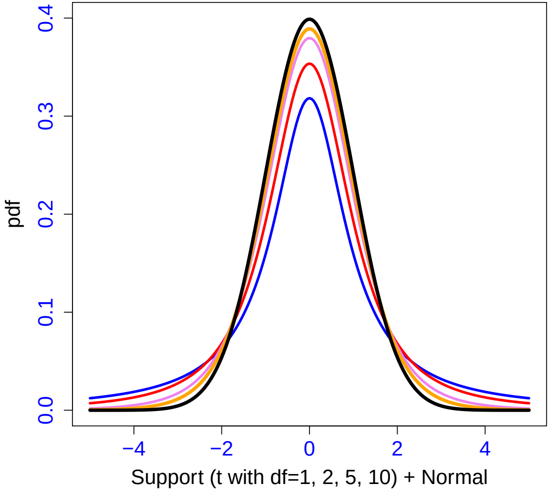
follows Student's t distribution with $n-1$ degrees of freedom

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

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

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


pdf
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
Support (t with df=1, 2, 5, 10) + Normal

R: t quantiles

```
##quantiles of a chi-square distribution
n=c(1,2,5,10)
alpha <- .05
c(qt(1-alpha/2,n),qnorm(1-alpha/2))

##results
[1] 12.71  4.30  2.57  2.23  1.96
```

Notes 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 of sleep 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
```

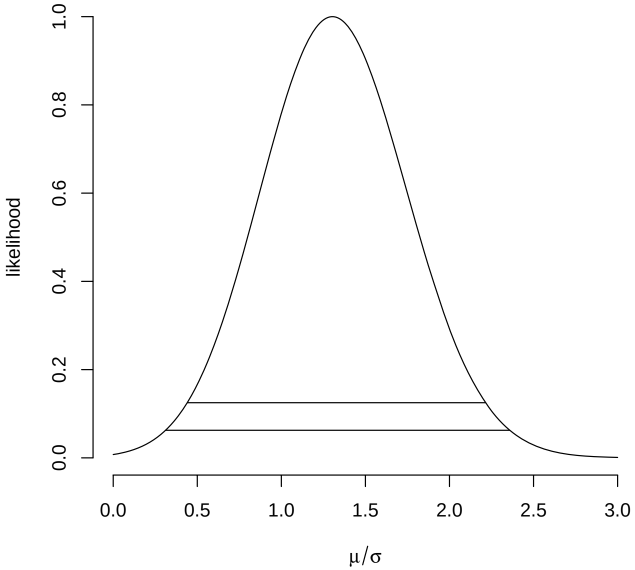
The non-central t distribution

- 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))
```


likelihood
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
μ/σ