# T Confidence Intervals

## Statistical Inference

**Brian Caffo, Jeff Leek, Roger Peng**
**Johns Hopkins Bloomberg School of Public Health**

# Confidence intervals

- In the previous, we discussed creating a confidence interval using the CLT
- In this lecture, we discuss some methods for small samples, notably 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 $\frac{(n - 1) S^2}{\sigma^2} \sim \chi^2_{n-1}$ 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 $\infty$
- The mean of the Chi-squared is its degrees of freedom
- The variance of the Chi-squared distribution is twice the degrees of freedom

# Confidence interval for the variance

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

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

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 $\sigma^2$

# Notes about this interval

- This interval relies heavily on the assumed normality
- Square-rooting the endpoints yields a CI for \sigma

# Example

Confidence interval for the standard deviation of sons' heights from Galton's data

```
library(UsingR)
data(father.son)
x <- father.son$sheight
s <- sd(x)
n <- length(x)
round(sqrt((n - 1) * s^2/qchisq(c(0.975, 0.025), n - 1)), 3)
```

```
## [1] 2.701 2.939
```

# 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 $\chi^2$ 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

- 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 $\mu$
- Let $t_{df,\alpha}$ be the $\alpha^{th}$ quantile of the t distribution with $df$ degrees of freedom
$$\begin{eqnarray*} & & 1 - \alpha \\ & = & P\left(-t_{n-1,1-\alpha/2} \leq \frac{\bar X - \mu}{S/\sqrt{n}} \leq t_{n-1,1-\alpha/2}\right) \\ \\ & = & P\left(\bar X - t_{n-1,1-\alpha/2} S / \sqrt{n} \leq \mu \leq \bar X + t_{n-1,1-\alpha/2}S /\sqrt{n}\right) \end{eqnarray*}$$
- 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.

# The data

```
data(sleep)
head(sleep)
```

```
##   extra group ID
## 1   0.7     1  1
## 2  -1.6     1  2
## 3  -0.2     1  3
## 4  -1.2     1  4
## 5  -0.1     1  5
## 6   3.4     1  6
```

```r
g1 <- sleep$extra[1:10]
g2 <- sleep$extra[11:20]
difference <- g2 - g1
mn <- mean(difference)
s <- sd(difference)
n <- 10
mn + c(-1, 1) * qt(0.975, n - 1) * s/sqrt(n)
```

```
## [1] 0.7001 2.4599
```

```r
t.test(difference)$conf.int
```

```
## [1] 0.7001 2.4599
## attr(,"conf.level")
## [1] 0.95
```