# Lecture 15 - Inference for Means

Sta 111

Colin Rundel

June 5, 2014

## Mean

- *Sample mean* ($\bar{x}$)

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- *Population mean* ($\mu$)

$$\mu = \frac{1}{N}(x_1 + x_2 + x_3 + \cdots + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- The sample mean is a *sample statistics*, or a *point estimate* of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population) it is usually a good guess.

## Variance

- *Sample variance* ($s^2$)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- *Population variance* ($\sigma^2$)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Note that the sample variance is divided by $n-1$ and not $n$. This is necessary for unbiasedness - $E(s^2) = \sigma^2$.

## Central Limit Theorem

**Central limit theorem**

The distribution of the sample mean is well approximated by a normal model:

$$\bar{x} \sim N\left(mean = \mu, \sigma = \frac{\sigma}{\sqrt{n}}\right)$$

when $n$ is large. If $\sigma$ is unknown, use $s$.

- As $n$ increases the standard deviation of $\bar{x}$ (often called the standard error) decreases.
- Large sample sizes yield more consistent sample means, hence the variability among the sample means should be lower.

## Confidence intervals

- A plausible range of values for the population parameter is called a *confidence interval*.
- Using only a point estimate to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we are more likely to miss. If we toss a net in that area, we have a better chance of catching the fish.

- If we report a point estimate, we probably will not hit the exact population parameter. If we report a range of plausible values – a confidence interval – we have a good shot at capturing the parameter.

## Confidence intervals and the CLT

We have a point estimate $\bar{x}$ for the population mean $\mu$, but we want to design a "net" to have a reasonable chance of capturing $\mu$.

From the CLT we know that we can think of $\bar{x}$ as a sample from $N(\mu,\ \sigma/\sqrt{n})$.

Therefore, 96% of observed $\bar{x}$'s should be within 2 SEs ($2\sigma/\sqrt{n}$) of $\mu$.

Clearly then for 96% of random samples from the population, $\mu$ must then be with in 2 SEs of $\bar{x}$.

Note that we are being very careful about the language here - the 96% here only applies to random samples in the abstract. Once we have actually taken a sample $\bar{x}$ will either be within 2 SEs or outside of 2 SEs of $\mu$.

## Example - Relationships

A sample of 50 Duke students were asked how many long term exclusive relationships they have had. The sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

The 96% confidence interval is defined as

$$point\ estimate \pm 2 \times SE$$

$$\bar{x} = 3.2 \qquad s = 1.74 \qquad SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.25$$

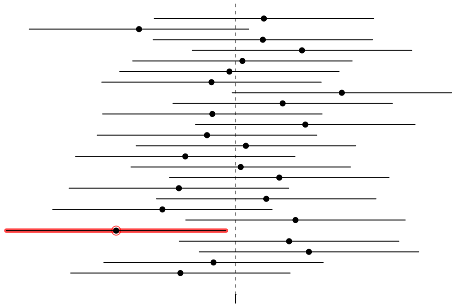
$$\begin{aligned} \bar{x} \pm 2 \times SE &= 3.2 \pm 2 \times 0.25 \\ &= (3.2 - 0.5, 3.2 + 0.5) \\ &= (3.15, 3.25) \end{aligned}$$

We are 96% confident that Duke students on average have been in between 3.15 and 3.25 exclusive relationships

## What does 96% confident mean?

- Suppose we took many samples and built a confidence interval from each sample using the equation *point estimate* $\pm 2 \times SE$.
- Then about 96% of those intervals would contain the true population mean ($\mu$).

- The figure on the left shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.



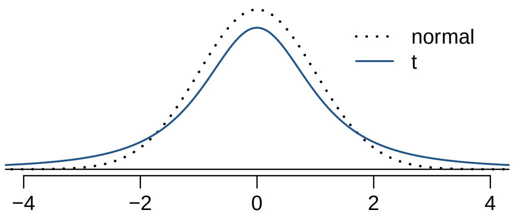
- It **does not** mean there is a 96% probability the CI contains the true value

## CLT and large samples

- As long as observations are independent a large sample ensures that the sample average will have a nearly normal distribution with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$.
- But when it comes to inference there is a problem, we almost never know $\sigma$.
- If $n$ is large enough then $s$ should be close to $\sigma$ just like $\bar{x}$ is close to $\mu$.
- What do we do when $n$ isn't large then? *Use a more conservative distribution*

## The *t* distribution

- When working with small samples, and the population standard deviation is unknown (this is almost always the case), the uncertainty of the standard error estimate is addressed by using a new distribution - the *t distribution*.
- This distribution is also bell shape, but its tails are *thicker* than the normal.
- These extra thick tails are helpful for resolving our problem with a less reliable estimate of the standard error (since $n$ is small)



## t distribution

Let $X \sim t(d)$, then

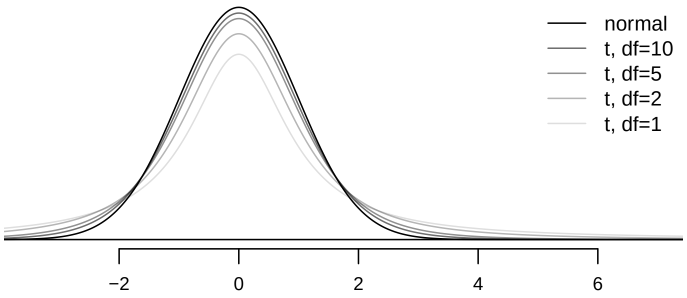
$$f(x|d) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{d\pi}\,\Gamma\left(\frac{d}{2}\right)}\left(1+\frac{x^2}{d}\right)^{-\frac{d+1}{2}}$$

$$\begin{aligned} Range(X) &= (-\infty, \infty) \\ E(X) &= 0 \\ Var(X) &= \begin{cases} \frac{d}{d-2} & \text{for } d > 2 \\ \infty & \text{otherwise} \end{cases} \end{aligned}$$

## Properties of the *t* distribution

The *t* distribution ...

- is always centered at zero, like the standard normal ($Z$) distribution.
- has a single parameter: *degrees of freedom* (*df*).



- as *df* increases the *t* distribution converges to the unit normal distribution.

## History of the $t$ distribution

First published by by William Gosset ...

- Oxford Graduate with a degree in Chemistry and Mathematics
- Hired as a brewer by the Guinness Brewery in 1899
- Spent 1906 - 1907 studying with Karl Pearson
- Published “The probable error of a mean” in 1908 under the pseudonym “Student”
- Much of his work was promoted by R.A. Fisher

## A more accurate interval

**Confidence interval, a general formula**

$$point\ estimate \pm CV \times SE$$

Conditions when the point estimate is $\bar{x}$:

1. *Independence*: Observations in the sample must be independent
   - random sample/assignment
   - $n <$ 10% of population
2. *Sample size / skew*: Reasonable sample size and distribution not extremely skewed

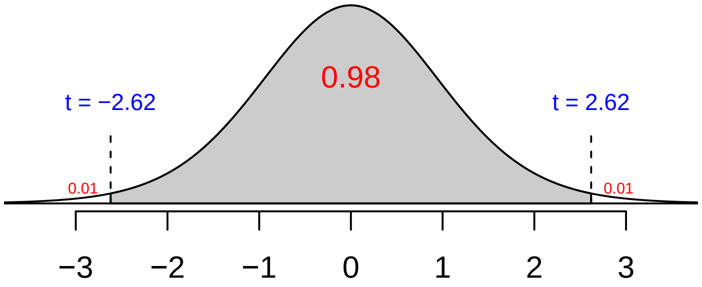
## Changing the confidence level

$$\bar{x} \pm t^\star \times SE$$

- In order to change the confidence level all we need to do is adjust $t^\star$ in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.
- Using the $t$ table it is possible to find the appropriate $t^\star$ for any confidence level and sample size (use $df = n - 1$).

## Example - Calculating $t^\star$

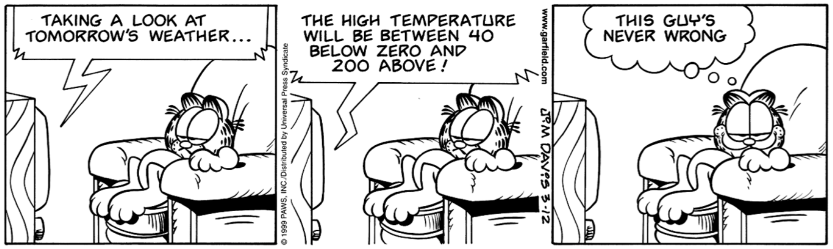
What is the appropriate value for $t^\star$ when calculating a 98% confidence interval for a sample of size 15?



## Width of an interval

If we want to be very certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

Can you see any drawbacks to using a wider interval?


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC./Distributed by Universal Press Syndicate
www.garfield.com
JIM DAVIS 3-12

## Example - Sample Size

Coca-Cola wants to estimate the per capita number of Coke products consumed each year in the United States, in order to properly forecast market demands they need their margin of error to be 5 items at the 95% confidence level. From previous years they know that $\sigma \approx 30$. How many people should they survey to achieve the desired accuracy? What if the requirement was at the 99% confidence level?

## Common Misconceptions

1. The confidence level of a confidence interval is the probability that the interval contains the true population parameter.

   *This is incorrect, CIs are part of the frequentist paradigm and as such the population parameter is fixed but unknown. Consequently, the probability any given CI contains the true value must be 0 or 1 (it does or does not).*

2. A narrower confidence interval is always better.

   *This is incorrect since the width is a function of both the confidence level and the standard error.*

3. A wider interval means less confidence.

   *This is incorrect since it is possible to make very precise statements with very little confidence.*

## Hypothesis testing framework

- We start with a *null hypothesis* ($H_0$) that represents the status quo.
- We develop an *alternative hypothesis* ($H_A$) that represents our research question (what we're testing for).
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or theoretical methods.
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

We'll formally introduce the hypothesis testing framework using an example on testing a claim about a population mean.

## Example - Grade inflation?

In 2001 the average GPA of students at Duke University was 3.37. Last semester Duke students in a Stats class were surveyed and ask for their current GPA. This survey had 147 respondents and yielded an average GPA of 3.56 with a standard deviation of 0.31.

Assuming that this sample is random and representative of all Duke students, do these data provide convincing evidence that the average GPA of Duke students has *changed* over the last decade?

## Setting the hypotheses

- The *parameter of interest* is the average GPA of current Duke students.
- There may be two explanations why our sample mean is higher than the average GPA from 2001.
  - The true population mean has changed.
  - The true population mean remained at 3.37, the difference between the true population mean and the sample mean is simply due to natural sampling variability.
- We start with the assumption that nothing has changed.

$$H_0 : \mu = 3.37$$

- We test the claim that average GPA has changed.

$$H_A : \mu \neq 3.37$$

## Making a decision - p-values

We would know like to make a decision about whether we think $H_0$ or $H_A$ is correct, to do this in a principled / quantitative way we calculate what is known as a *p-value*.

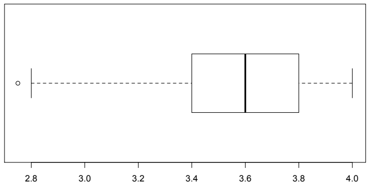
- The *p-value* is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis was true.
- If the p-value is *low* (lower than the significance level, $\alpha$, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject $H_0$*.
- If the p-value is *high* (higher than $\alpha$) we say that it is likely to observe the data even if the null hypothesis were true, and hence *do not reject $H_0$*.
- We never accept $H_0$ since we're not in the business of trying to prove it. We simply want to know if the data provide convincing evidence against $H_0$.

## Conditions for inference

In order to perform inference using this data set, we need to use the CLT and therefore we must make sure that the necessary conditions are satisfied:

1. *Independence:*
   - We have already assume this sample is random.
   - 147 < 10% of all current Duke students.

   ⇒ we are safe assuming that GPA of one student in this sample is independent of another.
2. *Sample size / skew:* The distribution appears to be slightly skewed (but not extremely) and $n$ is large so we can assume that the distribution of the sample means is nearly normal.



# Calculating the p-value

*p-value:* probability of observing data at least as favorable to $H_A$ as our current data set (a sample mean greater than 3.56 or less than 3.18), if in fact $H_0$ was true (the true population mean was 3.37).

$$
\begin{aligned}
&P(\bar{x} > 3.56 \text{ or } \bar{x} < 3.18 \mid \mu = 3.37)\\
&\quad = P(\bar{x} > 3.56 \mid \mu = 3.37) + P(\bar{x} < 3.18 \mid \mu = 3.37)\\
&\quad = P\left(t_{146} > \frac{3.56 - 3.37}{0.31/\sqrt{147}}\right) + P\left(t_{146} < \frac{3.18 - 3.37}{0.31/\sqrt{147}}\right)\\
&\quad = P\left(t_{146} > 7.43\right) + P\left(t_{146} < -7.43\right)\\
&\quad = 10^{-13} \approx 0
\end{aligned}
$$

# Drawing a Conclusion / Inference

$$p - value \approx 10^{-13}$$

- If the true average GPA Duke students applied to is 3.37, there is approximately a $10^{-11}$ % chance of observing a random sample of 147 Duke students with an average GPA of 3.56.
- This is a very low probability for us to think that a sample mean of 3.56 GPA is likely to happen simply by chance.
- Since p-value is *low* (lower than 5%) we *reject $H_0$*.
- The data provide convincing evidence that Duke students average GPA has changed since 2001.
- There is significant evidence that the difference between the null value of a 3.37 GPA and observed sample mean of 3.56 GPA is *not due to chance* or sampling variability.