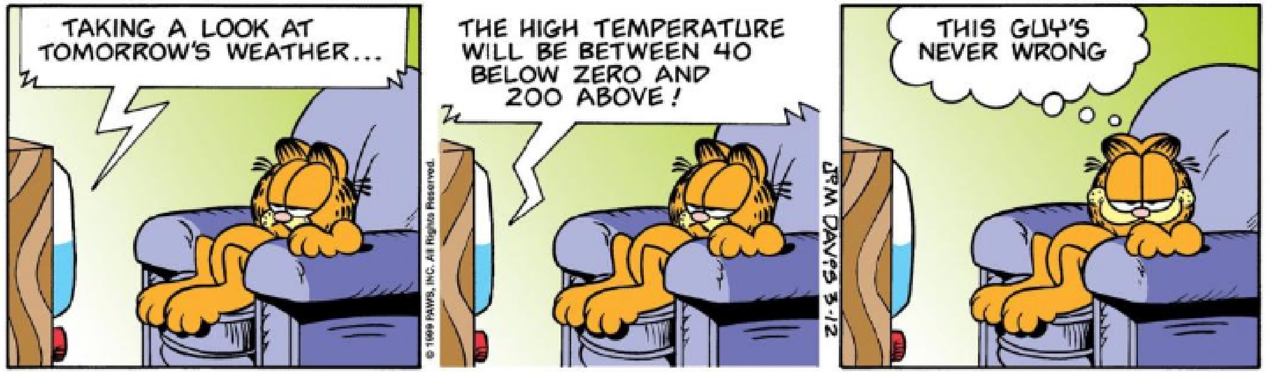
# *What are your best estimates? Are you close?*

**Estimation**:  *What are your best estimates (of the unknown parameters)?*

- LUEs and ***BLUE*** (minimum variance in the class of LUEs)
- So far:  No distributional assumptions (e.g. Normal Distribution).

**Inference**:  *Are you sure/close?*

- How close are your parameter estimates to the true underlying parameters?
- Generally:  Need to make distributional assumptions to proceed.
- Focus on the two main tools of inference
  - ***Confidence Intervals***:  Interval estimators
  - ***Hypothesis Testing***:  Can you (confidently) reject the *Null Hypothesis*?



# *The Tools of Inference*

***Confidence Intervals***

- Confidence intervals provide an interval estimate of the true parameter value.
- Some high percent (95%?) of the interval estimates generated in some fashion will in fact contain the true unknown parameter value.
- But is the true parameter contained in the specific interval you are looking at?  No idea!... but we do know that *some high percent (95%?) of the interval estimates generated this way*…

**II:  *Hypothesis Testing***

- Is the true (unknown) parameter zero?  (this will be the #1 Null Hypothesis for us)
- Your point estimate is very very far from zero.  But maybe you just have had a really wacky unrepresentative sample, and the true underlying parameter is in fact zero.  That is always a possibility… but is it probable?  How probable?
- Less than say 5% probable?
    - Then OK, we reject the Null Hypothesis that the true parameter is zero at the 5% *significance level*.
    - And we say that our estimate is *statistically significant* at that level.  We could be wrong… but that is not at all likely!  But for sure, it won't happen very often!

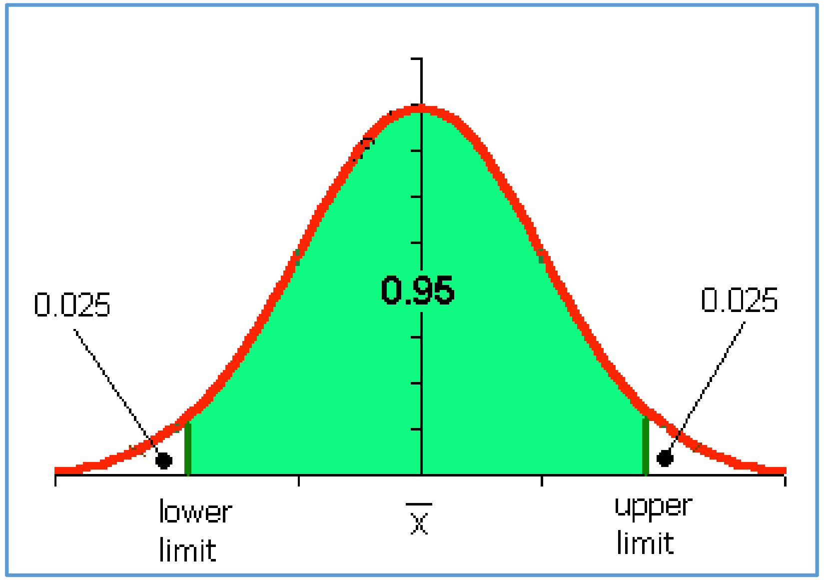
# Distributional Assumptions: *Generally required to do Inference*

**Not necessary for *Estimation***:

- We made no distributional assumptions in showing that the Sample Mean was ***BLUE***.

**But generally required to do Inference**

- We typically assume ***Normal distributions***. You can of course work with other distributions… but you have to start somewhere, and why not begin with an assumption of Normality?



# Estimating the Mean of the Distribution, *cont'd*

**Sampling and estimating**: Let's return to estimating the mean of the distribution of Y.

- You have an *iid* random sample $\{Y_1, Y_2, \ldots Y_n\}$ from the distribution of Y, which has unknown mean, $\mu$ .

- You are using the ***BLUE*** sample mean estimator, $\bar{Y} = \frac{1}{n}\sum Y_i$ , to estimate $\mu$ .

  - $E(\bar{Y}) = \mu$ and $Var(\bar{Y}) = \frac{\sigma^2}{n}$ , where $\sigma^2$ is the variance of Y.

  - Since we've made no distributional assumptions yet, the particular nature of the distribution of Y (or of $\bar{Y} = \frac{1}{n}\sum Y_i$ ) is as yet unknown.

# Distributional Assumption: *Assume a Normal distribution*

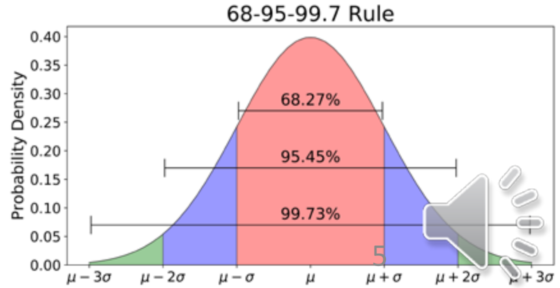
- Assume that Y is Normally distributed, so that $Y$ (and the *iid* $Y_i$'s) are all $\mathrm{N}(\mu,\sigma^2)$

- Since $\sum Y_i \sim \mathrm{N}(n\mu, n\sigma^2)$, $\bar{Y}$ is Normally distributed: $\bar{Y} = \frac{1}{n}\sum Y_i \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$

- The sample mean estimator, $\bar{Y}$, has mean $\mu$, variance $\frac{\sigma^2}{n}$, and standard deviation

  $sd\left(\bar{Y}\right) = \frac{\sigma}{\sqrt{n}}$.

- Put differently: $Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{Y}-\mu}{sd(\bar{Y})}$ has the standard Normal distribution, $Z \sim N(0,1)$.



# Confidence Intervals I: *Known variance*

- Start with an unrealistic example: The variance of *Y*, $\sigma^2$, is known.
- Here's a symmetric (confidence) interval estimator:
  - $\left[\bar{Y} - c\frac{\sigma}{\sqrt{n}}, \bar{Y} + c\frac{\sigma}{\sqrt{n}}\right]$, where $c \geq 0$ is some pre-specified *critical* value.
  - In words: This Confidence Interval is the Sample Mean, plus or minus c standard deviations of $\bar{Y}$, $sd\left(\bar{Y}\right)$: $CI = \left[est \pm c\ stdevs\right]$
- ***Observations***:
  - The only random component in this interval estimator is $\bar{Y}$, since n and the variance $\sigma^2$ are known, and c is pre-specified.
  - The interval will shift around depending on $\bar{Y}$, the *Sample Mean* around which it is centered, and with a constant width of $2c\frac{\sigma}{\sqrt{n}}$.

# Confidence Intervals I: *Known variance, cont'd*

- The probability that this random interval estimator contains the unknown mean $\mu$ is

$$prob\left(\mu \in \left[\bar{Y} - c\frac{\sigma}{\sqrt{n}}, \bar{Y} + c\frac{\sigma}{\sqrt{n}}\right]\right) = prob\left(\mu \in \left[\bar{Y} \pm c\frac{\sigma}{\sqrt{n}}\right]\right) = prob\left(-c \le \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \le c\right)$$

- Since $\dfrac{\bar{Y} - \mu}{\sigma / \sqrt{n}} = Z \sim N(0,1)$, this is $prob\left(-c \le N(0,1) \le c\right)$

- For a given level of confidence, this allows us to set the critical value c:

| critical val. c | p(-c<Z<c) | | p(-c<Z<c) | critical val. c |
|---|---|---|---|---|
| 1.5 | 86.6% | | 89% | 1.60 |
| 1.6 | 89.0% | | 90% | 1.64 |
| 1.7 | 91.1% | | 91% | 1.70 |
| 1.8 | 92.8% | | 92% | 1.75 |
| 1.9 | 94.3% | | 93% | 1.81 |
| 2 | 95.4% | | 94% | 1.88 |
| 2.1 | 96.4% | | 95% | 1.96 |
| 2.2 | 97.2% | | 96% | 2.05 |
| 2.3 | 97.9% | | 97% | 2.17 |
| 2.4 | 98.4% | | 98% | 2.33 |
| 2.5 | 98.8% | | 99% | 2.58 |

  - 90% Confidence Interval: $\left[\bar{Y} \pm 1.64\frac{\sigma}{\sqrt{n}}\right]$
  - 95% Confidence Interval: $\left[\bar{Y} \pm 1.96\frac{\sigma}{\sqrt{n}}\right]$

- **A *Good Rule of Thumb***: … the 95% confidence interval is the Sample Mean +/- about two standard deviations.

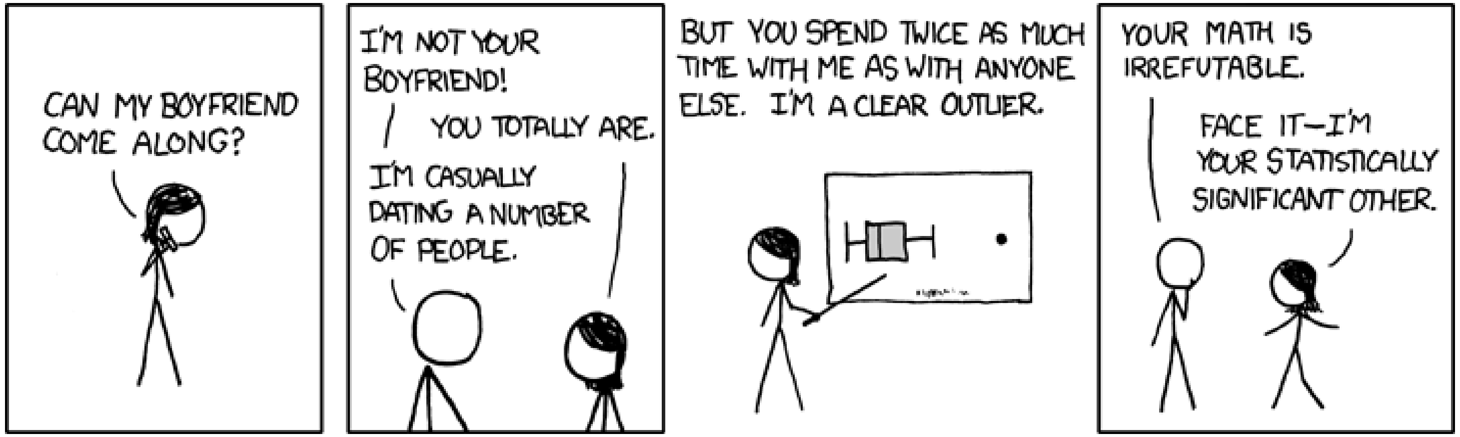
# Confidence Intervals II: *The variance of Y is now unknown*

- Since we don't know $\sigma^2$, we don't know $sd\left(\bar{Y}\right)=\frac{\sigma}{\sqrt{n}}$ … which makes it difficult to construct the Confidence Intervals we just considered.
- But the sample variance, $S_Y^2=\frac{1}{n-1}\sum(Y_i-\bar{Y})^2$, is an unbiased estimator of $\sigma^2$.
- And so $\frac{S_Y^2}{n}$ will be an unbiased estimator of $Var\left(\bar{Y}\right)=\frac{\sigma^2}{n}$.
- Taking the square root, we have $\frac{S_Y}{\sqrt{n}}$ as an estimator of the standard deviation of $\bar{Y}$.

# *The Standard Error*

- We call $se = se\left(\bar{Y}\right) = \frac{S_Y}{\sqrt{n}}$ the ***standard error*** (se) of the sample mean estimator… it's an estimate of $sd\left(\bar{Y}\right) = \frac{\sigma}{\sqrt{n}}$, the standard deviation of the (Sample Mean) estimator.

- ***Again***: The ***standard error*** of $\bar{Y}$, $se\left(\bar{Y}\right)$, provides an estimate of the ***standard deviation*** of $\bar{Y}$, $sd\left(\bar{Y}\right)$.



# *t Distributions and Standard Errors*

- If Y is normally distributed, then $\dfrac{\bar{Y}-\mu}{\sigma/\sqrt{n}} = \dfrac{\bar{Y}-\mu}{sd} \sim N(0,1)$.
- Now replace/estimate $\sigma$ with $S_Y$ … so we have the estimator: $\dfrac{\bar{Y}-\mu}{se\left(\bar{Y}\right)} = \dfrac{\bar{Y}-\mu}{S_Y/\sqrt{n}}$.
- This estimator will have a (Student's) t distribution with n-1 degrees of freedom. So $\dfrac{\bar{Y}-\mu}{S_Y/\sqrt{n}} \sim t_{n-1}$.
- The ***Student's t*** distribution was developed by William Sealy Gosset. in the early 1900's. At that time he was an employee (chemist and statistician) of Arthur Guinness & Son, the brewery in Dublin, Ireland.

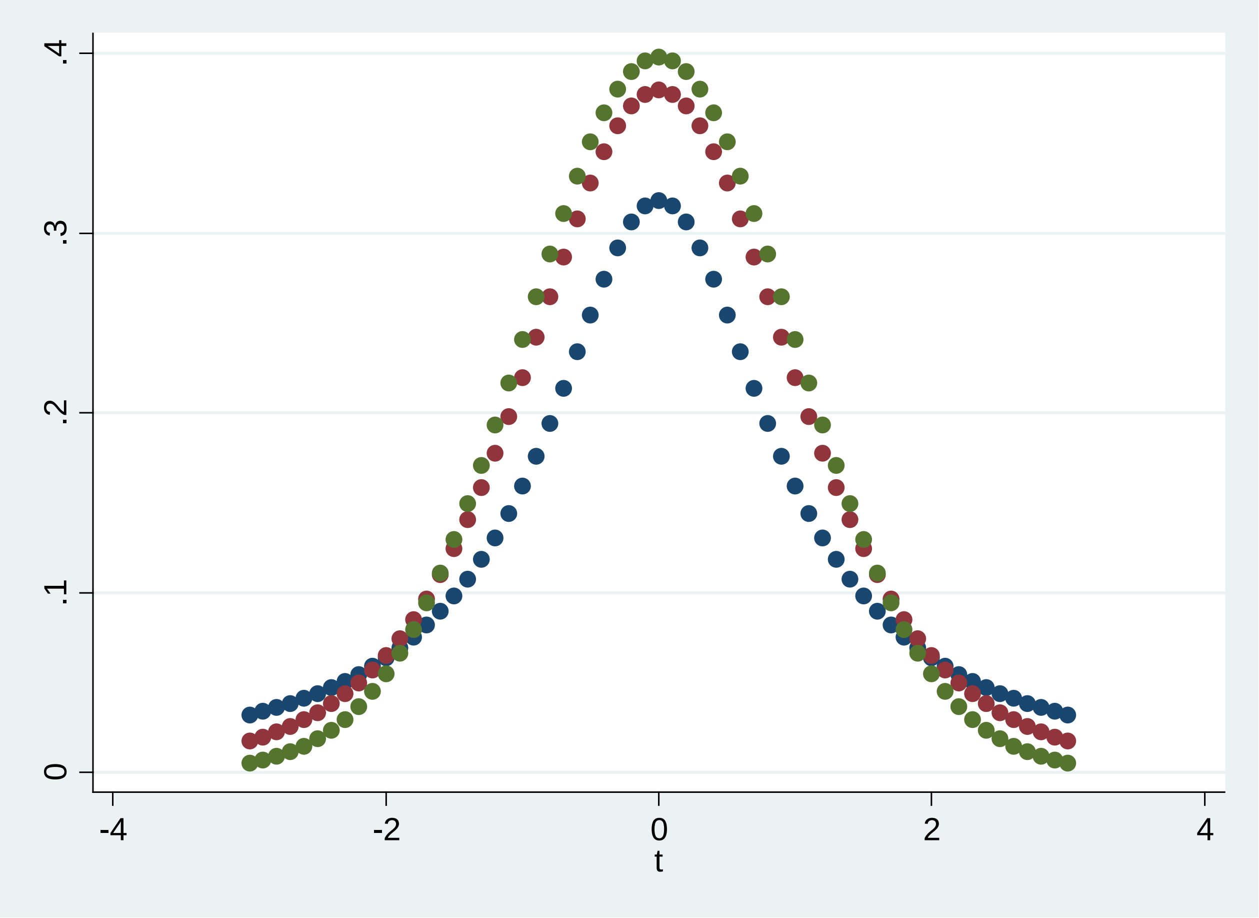
**William Sealy Gosset**

William Sealy Gosset (aka *Student*) in 1908 (age 32).

| | |
|---|---|
| **Born** | 13 June 1876<br>Canterbury, Kent, England |
| **Died** | 16 October 1937 (aged 61)<br>Beaconsfield, Buckinghamshire, England |
| **Other names** | Student |
| **Alma mater** | New College, Oxford |
| **Known for** | Student's t-distribution |

# *The t distribution looks a lot like the Normal distribution*

- Here are the density functions for three t distributions, with dof's = 1, 5 and 99. Notice that the density function is symmetric and bell-shaped, and centered around 0. As the *dofs* increase, probability shifts from the tails to the middle of the distribution. In the limit, and as *dofs* approach infinity, the t distribution approaches $N(0,1)$.



# The Cornerstone of Inference: *The t statistic*

**The Cornerstone of Inference:** ***The t statistic***

- $\dfrac{\bar{Y} - \mu}{S_Y / \sqrt{n}}$ is sometimes called the ***t-statistic***, and it drives inference (when using the Sample Mean to estimate the unknown mean, and the variance is unknown).
- ***Worth repeating!*** The t statistic drives inference!... and it has a t distribution with n-1 dofs.

# *Critical Values* (with unknown variance)

- Here's a symmetric (confidence) interval estimator: $\left[\bar{Y} - c\frac{S_Y}{\sqrt{n}}, \bar{Y} + c\frac{S_Y}{\sqrt{n}}\right]$ or $\left[\bar{Y} \pm c\frac{S_Y}{\sqrt{n}}\right]$

  $c \geq 0$ is a "critical" value, determined using the t distribution with n-1 dofs.

- As before, and after some algebra ($t_{n-1}$ is a t distribution with n-1 dofs):

$$prob\left(\mu \in \left[\bar{Y} - c\frac{S_Y}{\sqrt{n}}, \bar{Y} + c\frac{S_Y}{\sqrt{n}}\right]\right) = prob\left(-c \leq \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \leq c\right) = prob\left(-c \leq t_{n-1} \leq c\right).$$

- Given level of confidence and #dofs, we set the critical value c using the $t_{n-1}$ distribution.

# The Cornerstone of Inference: *The t statistic*, *cont'd*

- Before we had: $CI = \left[ est \pm c \; stdevs \right]$ … now we have: $CI = \left[ est \pm c \; sterrs \right]$
- 90% Confidence Interval
  - #dofs = 25: $\left[ \bar{Y} \pm 1.71 \frac{S_Y}{\sqrt{n}} \right]$
  - #dofs = infinite, N(0,1): $\left[ \bar{Y} \pm 1.64 \frac{S_Y}{\sqrt{n}} \right]$
- 95% Confidence Interval
  - #dofs = 25: $\left[ \bar{Y} \pm 2.06 \frac{S_Y}{\sqrt{n}} \right]$
  - #dofs = infinite, N(0,1): $\left[ \bar{Y} \pm 1.96 \frac{S_Y}{\sqrt{n}} \right]$

| dofs | 90.0% | 92.5% | 95.0% | 97.5% | 99.0% |
|---|---|---|---|---|---|
| 5 | 2.02 | 2.24 | 2.57 | 3.16 | 4.03 |
| 10 | 1.81 | 1.99 | 2.23 | 2.63 | 3.17 |
| 15 | 1.75 | 1.91 | 2.13 | 2.49 | 2.95 |
| 20 | 1.72 | 1.88 | 2.09 | 2.42 | 2.85 |
| 25 | 1.71 | 1.86 | 2.06 | 2.38 | 2.79 |
| 30 | 1.70 | 1.84 | 2.04 | 2.36 | 2.75 |
| 50 | 1.68 | 1.82 | 2.01 | 2.31 | 2.68 |
| 75 | 1.67 | 1.81 | 1.99 | 2.29 | 2.64 |
| 100 | 1.66 | 1.80 | 1.98 | 2.28 | 2.63 |
| | | | | | |
| ***infinite*** | ***1.64*** | ***1.78*** | ***1.96*** | ***2.24*** | ***2.58*** |

- ***Good Rule of Thumb***: For a 95% confidence interval, use the sample mean $\pm$ a couple standard errors. We'll use the same rule when you get to SLR and MLR analyses.