Notes

Week 3

Module 2 Week 3A

Point Estimation and Confidence Intervals

- I.e. how to quantify uncertainty of estimates?

Point Estimation

- E.g.:
  - What is the mean price of a 4-star hotel room in Washington DC?
    - Gather data (e.g. internet search of prices) -> calculate average price
      - This is a *point estimate* of a population mean
  - How often are flights delayed?
    - Gather data (e.g. record the flights that are delayed) à divide by the total number of flights.
      - This is a *point estimate* of a population proportion
- **Inferential Statistics** - using a sample of data to make predictions about population parameters of interest

Confidence intervals

- The *Point Estimate* - unlikely to be exactly equal to the true value of the parameter of interest, we hope it is close though.
  - A **confidence interval** - quantifies this uncertainty.
- A *confidence interval*, like other sample statistics, is a *random variable*, it is not fixed. (*Population Parameter* is fixed).
- E.g. Back to the hotel room example. A (random) sample of data is collected. The sample mean is calculated. We would like to know the population *standard deviation* (σ), but it is often unknown. The sample standard deviation (s) can be calculated. With this information a *confidence interval* for the estimated mean can be calculated. It changes with the *Random Sample*.
- A *confidence interval* provides a range of values it is reasonable for the *population mean (parameter)* to fall in.
  - There is no guarantee that the interval contains the true value of the parameter.
  - We can make probability statements about how likely it is.
- Recall that $\bar{X} \sim (\mu, \frac{\sigma^2}{n})$
  - *About 95% of all observations fall within 2 standard deviations of the mean.* I.e. about 95% of all sample means must fall within 2 *standard deviations* of μ
- E.g. suppose $\bar{X} = 290$ and $\sigma = 200$ (i.e. the *population SD* is known here) and $n = 100$
  - *Standard deviation* for $\bar{X} = \frac{\sigma}{\sqrt{n}} = \frac{200}{\sqrt{100}} = 20$
  - 2 standard deviations $= 40$
  - 95% *confidence interval* -> $290 \pm 40 = (250, 330)$

        - *We are 95% confident the population mean falls in this interval* (In reality it might not be, but probability that it doesn't is only 5%).
- *Confidence Intervals* form:
    - *Point Estimate* $\pm$ *Margin of Error*
        - *Margin of Error* depends on *level of confidence* and the *standard deviation of the estimate* (i.e. *standard error*)
            - **Error Bound** (for the population mean) - *Margin of Error* if the *Population Standard Deviation* is known.
        - **Standard Error** - *Standard Deviation* of a statistic.
- *Interpretation*: The **Confidence Level** is *the percentage of times the interval contains the true parameter value in repeated random samples*.
- $\mathbb{P}(\text{"CI does NOT contain the true value in repeated random samples"}) = \alpha = (1\text{–level of confidence}) = \text{level of significance}$
    - **Confidence Level** = $1 - \alpha$
- When σ is known, the *standard normal distribution* is used to calculate the confidence interval.
    - We need to find the value of z that puts the confidence level in the middle of the distribution and the level of significance in the tails. Each tail contains the area $= \alpha/2$
    - For a 95% confidence interval: $z = \pm 1.96$ (via software or the Z-Table).
- E.g.: what is the average amount of time spent commuting to work? Suppose in a random sample of $n = 36$ commuters the average commute time $\bar{X} = 27$ minutes and assume SD is known: $\sigma = 12$.
    - Recall: *Standard Deviation* for a *Sample Mean* statistic $\bar{X}$ is $SE = \frac{\sigma}{\sqrt{n}}$
    - 95% confidence interval $= 27 \pm 1.96 \frac{12}{\sqrt{36}} = 27 \pm 1.96 * 2 = (23.08, 30.92)$
    - 90% confidence interval $= 27 \pm 1.64 \frac{12}{\sqrt{36}} = 27 \pm 1.64 * 2 = (23.72, 30.28)$
- Problem: usually, we don't know the *Population Standard Deviation* σ
- When *Standard Deviation* σ is unknown:
    - If n is large enough (≥ 30): can be estimated with s, i.e. using the sample.
    - If the sample size is small: *Normal Distribution* cannot be used b/c the actual distribution depends on the sample size (discovery by William "the Student" Gosset) => the *Student's t-distribution* is used to calculate the confidence interval instead.
        - Until the 1970's, if $n > 30$, *Normal Distribution* is used. But today: just the the *Student T-Distribution* (thank your computer).
        - If you draw a simple random sample of size n from an approximate *normal distribution* with mean μ and unknown standard deviation and calculate *t-scores* ($\frac{\bar{X}-\mu}{s/\sqrt{n}}$), the t-scores follow a *Student's t-distribution* with $(n-1)$ *degrees of freedom* (*df*).
            - *T-Score Interpretation*: just like the *Z-Score*, measures how far $\bar{X}$ is from μ.
            - $(n-1)$ interpretation: from having to estimate the *Standard Deviation*.
        - *Student's t-distribution* is *symmetric*, has a mean of 0, but with *fatter tails* than the *standard normal distribution*. It *converges into the standard normal distribution as n increases*.

Proportions

- In addition to means, we are often interested in proportions.
    - Examples:
        - Proportion who vote for a candidate

        - Proportion of workers who are unemployed
        - Proportion of households with internet access
- Calculating confidence intervals for proportions is done in essentially the same way as for means, only the formulas differ.
- For a proportion, the underlying variable follows a *binomial distribution*.
- To calculate a proportion, take X (i.e. the number of “successes”) and divide by n (i.e. the number of trials)
- When n is large, we can approximate the *binomial distribution* with a *normal distribution* with mean np and *standard deviation* $\sqrt{npq}$.
- Dividing by n gives a proportion with mean p and standard deviation $\frac{\sqrt{pq}}{\sqrt{n}}$
    - The proportions follow a *normal distribution*
- *Confidence Intervals for Proportions*
    - Proportions follow an approximate normal distribution
        - Need to calculate z-scores to calculate confidence interval:
            - $z = \frac{p'-p}{\sqrt{pq/n}}$
    - *Confidence interval* $= p' \pm z * \text{standard error}$
    - E.g.: let $X = \text{number of households with internet access}$. Suppose in a sample of 400 households, 280 have internet access. I.e. $X \sim Bin(n = 400, p = \frac{280}{400})$
        - Estimate for the proportion: $p' = (280/400) = 0.70$
        - $z = 1.96$ because we're using the normal approximation with a 95% Confidence
        - 95% confidence interval $= 0.70 \pm 1.96(\frac{\sqrt{1.70*0.30}}{\sqrt{400}}) = 0.70 \pm 0.045 = (0.655, 0.745)$;
    - **“Plus four rule”**:
        - Should be used when:
            - We want confidence of $\geq 90\%$
              AND
            - We have $n \geq 10$ observations.
        - We simply pretend that we have four additional observations. Two of these observations are successes and two are failures. The new sample size, then, is $n + 4$, and the new count of successes is $x + 2$.