1 Review Problems

1.1 Example

- The Indian False Vampire bat consumes frogs. Assume the consumption time has a normal distribution with mean 27 minutes and known standard deviation 8 minutes. (Of course these population values are not known in reality).

  – For a single frog, what is the probability that consumption time is greater than 25 minutes?

  Let X be the consumption time of this type of bat. We have population information here, so $X \sim N(27, 8)$

  $$P(X > 25) = P(\frac{X - 27}{8} > \frac{25 - 27}{8}) = P(Z > -.25) \approx 0.6 \quad (1)$$

  Without a normal table you should be able to come up with a range of between 50-84 %, with your answer being closer to 50% than 80%.

  – *I will change the name of the bat so that in this case we don't know the population values and must estimate them.* The eating habits of 12 echolocating bats were examined For these 12 bats the average time to consume a frog was calculated to be 21.9 minutes, with sample standard deviation 8 minutes.

  * What is the probability that the mean consumption time is greater than 25 minutes?
    Note that above the standard deviation is not known but must be calculated.

    $$P(\ \overline{X} > 25) = P(\frac{\overline{X} - 25}{8/\sqrt{12}} > \frac{21.9 - 25}{8/\sqrt{12}}) = P(t_{11} > -1.34) \approx 0.896 \quad (2)$$

  * Construct and interpret a 95% confidence interval for the mean suppertime of a vampire bat whose meal consists of a frog. Do this for two cases:
    · σ is known to be 7.7 minutes
    $21.9 \pm z_{.975} 7.7/\sqrt{12}$
    · σ is estimated from the data to be 7.7 minutes
    $21.9 \pm t_{.975,11} 7.7/\sqrt{12}$

1.2 The Normal Distribution (Notes from STA240)

Density of $N(\mu, \sigma^2)$:

$$f(Y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{Y-\mu}{\sigma})^2} \quad (3)$$

Completely defined by the population parameters: μ (mean) and σ^2 (variance)

1.3 Standard Normal Distribution

- For $Y \sim N(\mu, \sigma^2)$, $Z = \frac{Y-\mu}{\sigma} \sim N(0, 1)$. Z is called the standard normal variate.
- Table of standard normal distribution: Table A.1, p. 708, R&S
- Let $W \sim N(0, 1)$. Find z such that $P(-z \leq W \leq z) = 1 - \alpha$. This z is denoted $z_{1-\alpha/2}$ and is called a *z-value*. For example, let $\alpha = .05$. From A.1, $z_{0.025} = 1.96$.

1.4 Why use the normal distribution?

- Many variables follow an approximately normal distribution
  – scores on tests taken by many people
  – repeated careful measurements of the same quantity
  – characteristics of biological populations (heights, weights)
- If a variable is non-normal, there is often a transformation to normality (3.5, R&S)
- *Central Limit Theorem*: Draw a simple random sample of size n from any population with mean μ and standard deviation σ. For n large, the sampling distribution of the sample mean $\bar{Y}$ is approximately normal.

1.5 Central Limit Theorem: Sampling distribution of the sample average

Central Limit Theorem (CLT): For population with mean μ and standard deviation σ,

1. the mean value of the collection of all possible sample means will equal the mean of the original population;
2. the standard deviation of the collection of all possible means of samples of a given size is $\frac{\sigma}{\sqrt{n}}$;
3. the distribution of sample means will be approximately normal *regardless of the distribution of values in the original population from which the samples were taken.*

$$\bar{Y} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \tag{4}$$

1.6 Application of the CLT: confidence limits for μ when σ is known.

- A 100(1-α)% confidence interval for μ: $\bar{Y} \pm z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$
- *Interpretation*: If we repeatedly drawing a sample of size n and form the interval each time, 95% of these *random* intervals will contain the true fixed value μ.
- *An example*: Crop researchers plant 15 plots with a new variety of corn. The mean yield in bushels per acre is 130. Assume σ=10 bushels per acre. Find the 90% confidence interval for the mean yield μ for this variety of corn. (Ans: [125.75,134.25])
- Sample size issues
  - In the above example, how many plots of corn are needed to estimate the average yield to within $\pm$ 2 bushels per acre with probability 95%?
  - Ans: Solve $2 = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ for n. $n = 96$. *As the confidence interval on the mean becomes narrower, what happens to n? What is happening to SE(mean) as n increases?*

1.7 Application of the CLT: hypothesis test for μ when σ is known.

- Two-sided test

Null hypothesis:	$H_o : \mu = \mu_o$
Alt. hypothesis:	$H_A : \mu =\neq \mu_0$

- Test statistic:

$$Z = \frac{\bar{X} - \mu_o}{\sigma/\sqrt{n}} \tag{5}$$

- Could the observed value of Z have come from a $N(0,1)$ distribution? Use of a "rejection rule" or a "p-value".
- One-sided vs. two-sided tests
- What is the relationship between confidence intervals and rejection regions for hypothesis tests?

1.8 σ unknown: Standard error of the sample average

- Consider a simple random sample (SRS) of size n (n large). Sample mean $\bar{Y} = \frac{\sum Y_i}{n}$ and standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$ are estimates of the mean and standard deviation of the population from which the sample was drawn.
- How accurate is our estimate of the mean? How can we measure the variability of possible values of means of samples of n from N?
  - By the CLT, under repeated sampling the variation in the values of $\bar{Y}$ is $\frac{\sigma}{\sqrt{n}}$.
  - Estimate this variation using $SE(\bar{Y}) = \frac{s}{\sqrt{n}}$, with degrees of freedom $df = n - 1$.
  - *degrees of freedom*: amount of information used to measure variability (estimation of s).
  - $SE(\bar{Y})$ decreases as n increases.

1.9 The t-distribution

- In most applications, σ is unknown, and is estimated by s. This *approximation* is usually satisfactory for $n \geq 30$. For the studies considered in the *Sleuth*, you will use the t-distribution for most cases.
- *When we have to estimate s, the quantity $\frac{\bar{Y}-\mu}{s/\sqrt{n}}$ is not normally distributed.*
- The quantity $\frac{\bar{Y}-\mu}{s/\sqrt{n}}$ has a t distribution with $n-1$ degrees of freedom.

$$\frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{n-1} \tag{6}$$

  In chapter 3, we'll discuss the applicability of the t distribution and two-sample t-tests for different sample sizes, standard deviations, and departures from independence.
- For n large, the t distribution with n degrees of freedom approaches the standard normal distribution.
- Percentiles of the t distribution: Table A.2, p. 710, R&S.

1.10 Applications

- A 100(1-α) confidence interval for the mean μ when σ is unknown and must be estimated.

$$\bar{Y} \pm t_{1-\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right) \tag{7}$$

- *An example*: In crop example above, assume σ unknown, but $s = 10$ bushels/acre. Ans: 90% CI: [125.45, 134.55] *Why and in what way is this interval different from the case when σ is known?*

- Hypothesis test for the mean μ when σ is unknown
  - Test statistic: $T = \frac{\bar{X}-\mu_o}{s/\sqrt{n}}$ under $H_o : \mu = \mu_o$
  - Under H_o, $T \sim t_{n-1}$
  - Two-sided test: $H_A : \mu \neq \mu_o$
  - Rejection region: Reject at level α if $|T| > t_{1-\alpha/2,n-1}$

A specific statistical test has an *observed significance level*, or ***p-value***. This is the probability of observing what you observed or something more extreme assuming the null hypothesis is true. The null hypothesis will be rejected only if the p-value is *less* than the fixed significance level α. An example follows.

1.11 An Example

The EPA sets a limit of 5 ppm on PCB in water. A major manufacturing firm producing PCB for electrical insulation discharges small amounts from the plant. The company management, attempting to control PCB discharge levels, has given instructions to halt production if the mean amount of PCB in the effluent exceeds 3 ppm.

A random sampling of 30 water specimens produces the following statistics: $\bar{Y} = 3.1$ ppm; $s = 0.5$ ppm.

- Do these statistics provide sufficient evidence for the managers to halt the production process? That is, the managers wish to claim that the mean PCB amount is less than 3 ppm. Interpret the p-value. As a plant manager, what value might you use for α (the required significance level)?

- As a scientist who has studied the adverse impact of PCB levels on the local ecosystem, what test and what α might you use? Would your hypotheses change?