Statistics Review – Part 2

Sample mean

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (3.1)$$

- From heights example: $\bar{y} = 174.1$, $\mu_y = 176.8$
- The sample (the y_i) were drawn randomly
- y is random $\rightarrow$ $\bar{y}$ is a random variable!
- $\bar{y}$ has a *sampling distribution* (probability function)
- The mean and variance of $\bar{y}$ will determine if $\bar{y}$ is:

- Unbiased
- Efficient
- Consistent

$$
\begin{aligned}
\mathrm{E}\left[\bar{y}\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n}\mathrm{E}\left[y_1 + y_2 + \cdots + y_n\right] \\
&= \frac{1}{n}\left(\mathrm{E}\left[y_1\right] + \mathrm{E}\left[y_2\right] + \cdots + \mathrm{E}\left[y_n\right]\right) \\
&= \frac{1}{n}\left(\mu_y + \mu_y + \cdots + \mu_y\right) \\
&= \frac{n\mu_y}{n} = \mu_y
\end{aligned}
\tag{3.2}
$$

3.2.4 Efficiency

An estimator is efficient if it has the smallest variance among all other potential estimators (for us, potential = linear, unbiased)

Need to get the variance of $\bar{y}$.

$$
\begin{aligned}
\mathrm{Var}[\bar{y}] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n^2}\mathrm{Var}\left[y_1 + y_2 + \cdots + y_n\right] \\
&= \frac{1}{n^2}\left(\mathrm{Var}[y_1] + \mathrm{Var}[y_2] + \cdots + \mathrm{Var}[y_n]\right) \\
&= \frac{1}{n}\left(\sigma_y^2 + \sigma_y^2 + \cdots + \sigma_y^2\right) \\
&= \frac{n\sigma_y^2}{n^2} = \frac{\sigma_y^2}{n}
\end{aligned} \tag{3.3}
$$

- Gauss-Markov theorem proves this is minimum variance
- We'll also need this to prove consistency, and for hyp. testing

3.2.5 Consistency

Suppose we had a lot of information. ($n \to \infty$)

What value should we get for our estimator?

How would state this mathematically?

Q) Prove that the sample mean is a consistent estimator for the population mean.

Q) Define the terms unbiasedness, efficiency, and consistency.

3.3 Hypothesis tests (known σ_y^2)

$$\begin{aligned} H_0 &: \mu_y = \mu_{y,0} \\ H_A &: \mu_y \neq \mu_{y,0} \end{aligned} \tag{3.4}$$

- Estimate μ_y (using $\bar{y}$ for example)
- See if $\bar{y}$ appears "close" to $\mu_{y,0}$
  - Remember, $\bar{y}$ is random! (and Normal)
- If it's close → fail to reject
- If it's far → reject

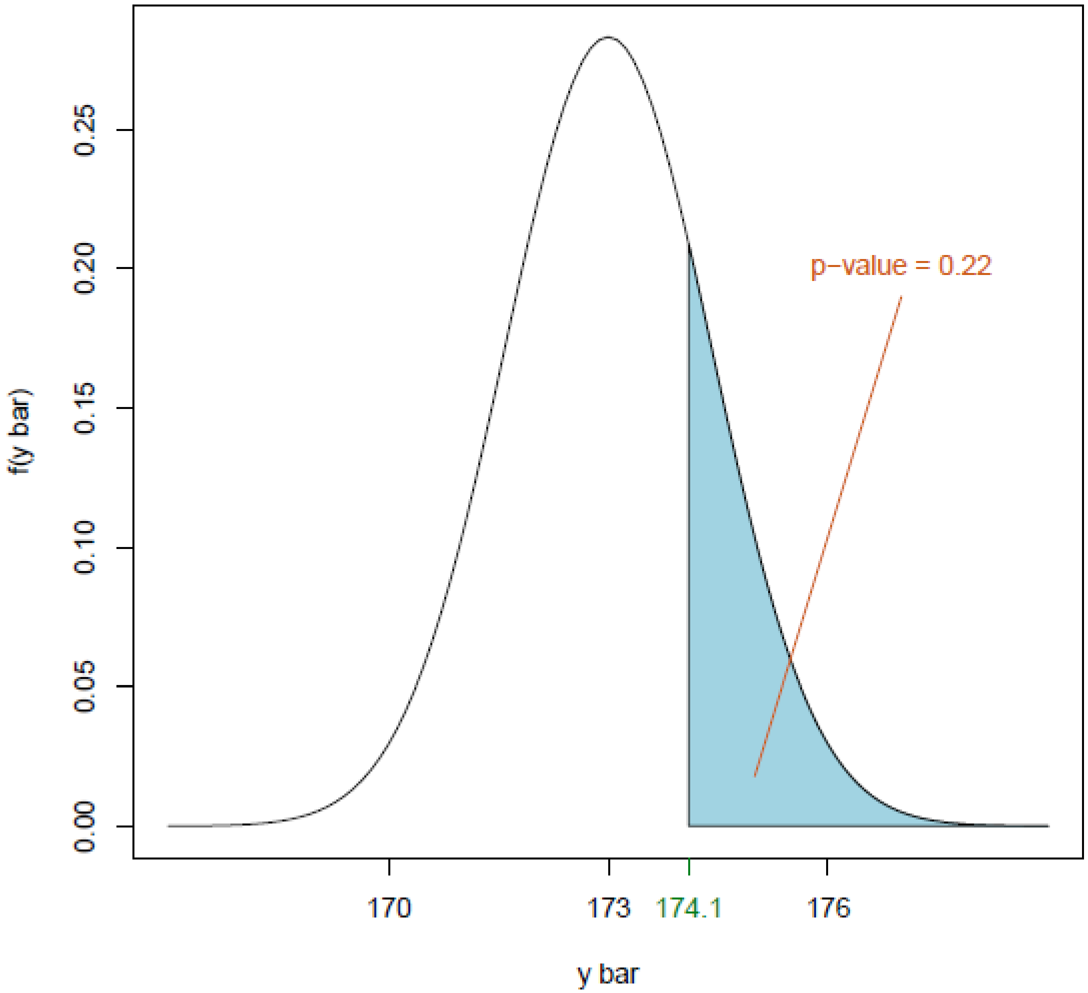
Example:

- Hypothesize that mean height of a U of M student is 173cm

$$H_0 : \mu_y = 173$$
$$H_A : \mu_y \neq 173 \qquad (3.5)$$

- Collect a sample: $y = \{173.9, 171.7, \ldots, 172.0\}$
- Calculate $\bar{y} = 174.1$
- Suppose (very unrealistically) that $\sigma_y^2 = 39.7$
- What now?

Figure 3.2: Normal distribution with $\mu = 173$ and $\sigma^2 = {}^{39.7}/_{20}$. Shaded area is the probability that the normal variable is greater than 174.1.



The p-value for the above test is 0.44. How to interpret this?

3.3.1 Significance of a test

3.3.2 Type I error

3.3.3 Type II error (and power)

3.3.4 Test statistics

- Just a more convenient way of getting the p-value for the test
- Each hypothesis test would present us with a new normal curve that we would have to draw, and calculate a new area (see fig. 3.2)
- Instead: *standardize*
- This gives us *one curve for all testing problems* (the standard normal curve)
- Calculate a bunch of areas under the curve, and tabulate them
- Not an issue with modern computers, but this is still the way we do things

3.3.5 Critical values

3.3.6 Confidence intervals

What is the probability that our z statistic will be within a certain interval, if the null hypothesis is true? For example, what is the following probability?

$$\Pr\left(-1.96 \leq z \leq 1.96\right)? \tag{3.12}$$

$$\Pr\left(-1.96 \leq \frac{\bar{y} - \mu_{y,0}}{\sqrt{\sigma_y^2/n}} \leq 1.96\right) = 0.95 \tag{3.13}$$

Finally, we solve equation 3.13 so that the null hypothesis $\mu_{y,0}$ is in the middle of the probability statement:

$$\Pr\left(\bar{y} - 1.96 \times \sqrt{\frac{\sigma_y^2}{n}} \leq \; \mu_{y,0} \; \leq \bar{y} + 1.96 \times \sqrt{\frac{\sigma_y^2}{n}}\right) = 0.95 \tag{3.14}$$