Statistics Review – Part 3

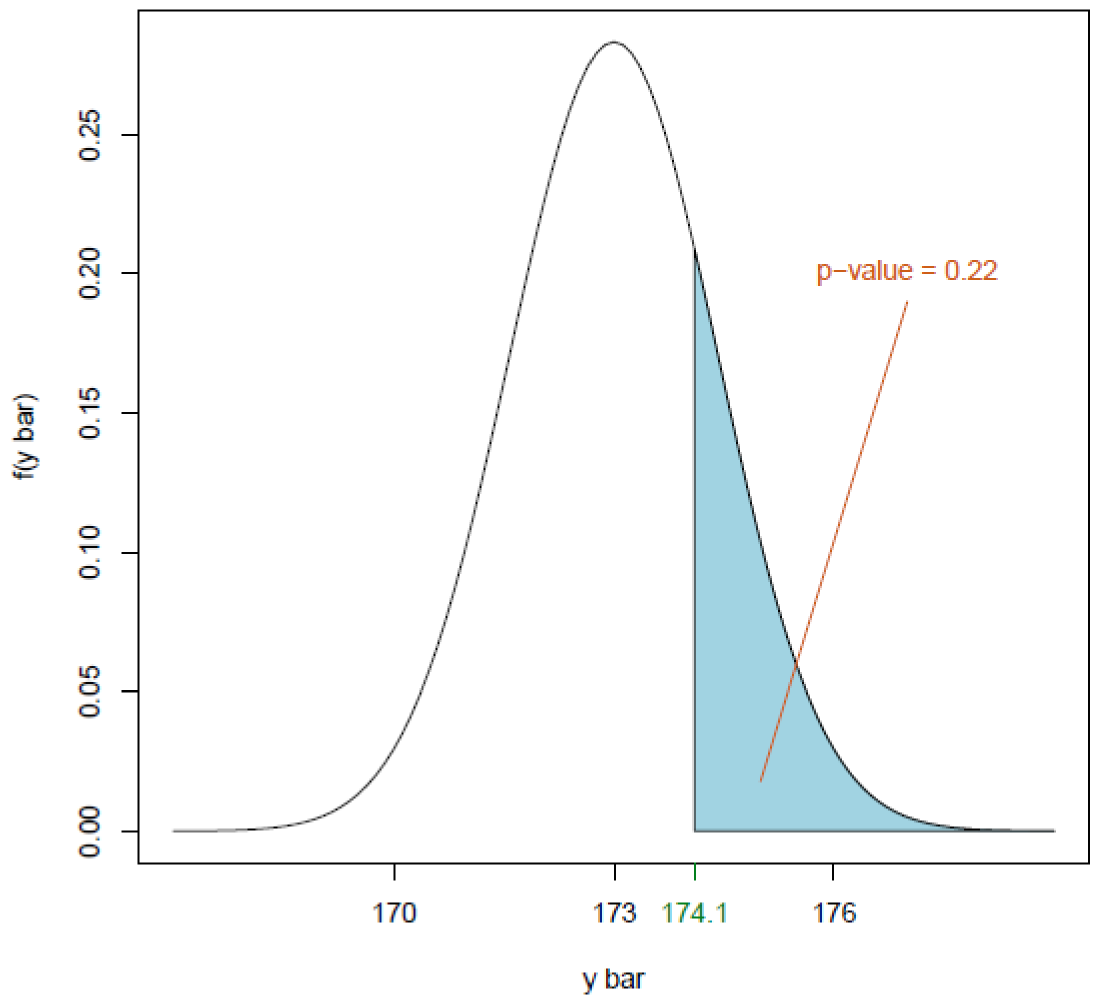
3.3 Hypothesis tests (known σ_y^2)

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

$$\begin{aligned} H_0 &: \mu_y = 173 \\ H_A &: \mu_y \neq 173 \end{aligned} \qquad (3.5)$$

- Collected a sample: $y = \{173.9, 171.7, \ldots, 172.0\}$
- Calculated $\bar{y} = 174.1$
- Supposed (very unrealistically) that we knew $\sigma_y^2 = 39.7$

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.



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
- How to get a z test statistic?
- Do a z test for our heights example.

Table 3.2: Area under the standard normal curve, to the right of z.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002

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:

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

3.4 Hypothesis Tests (unknown σ_y^2)

- Much more realistically, σ_y^2 (variance of y) will be unknown.
- Recall that: $Var[y] = {\sigma_y^2}/{n}$
- $z = \frac{\bar{y} - \mu_{y,0}}{s.e.(\bar{y})} = \frac{\bar{y} - \mu_{y,0}}{\sqrt{\sigma_y^2 / n}}$
- So, we need to estimate σ_y^2 in order to perform hypothesis tests.

3.4.1 Estimating σ_y^2

- A "natural" estimator:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3.15}$$

- Is this a good estimator? Why or why not?
- A better estimator:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (3.17)$$

- Degrees-of-freedom correction

So:

$$\text{Estimated variance of } \bar{y} = \frac{s_y^2}{n}$$

We can implement hypothesis testing by replacing the unknown σ_y^2 with its estimator s_y^2. The z test statistic now becomes:

$$\frac{\bar{y} - \mu_{y,0}}{\sqrt{s_y^2 / n}} = t$$

Note: for large *n*, the *t* test is equivalent to the *z* test

Chapter 2 Review Question

2. Let X be a random variable, where $X = 1$ with probability 0.5, and $X = -1$ with probability 0.5. Let Y be a random variable, where $Y = 0$ if $X = -1$, and if $X = 1$, $Y = 1$ with probability 0.5, and $Y = -1$ with probability 0.5. (a) What is the $\mathrm{Cov}(X, Y)$? (b) Are X and Y independent?

Chapter 3 Review Questions

3. Assume that $y_i \sim (\mu_y, \sigma_y^2)$, and that y_i is i.i.d. Let $\tilde{\mu}_y = \frac{y_1 + y_n}{2}$. Is $\tilde{\mu}_y$ an unbiased estimator for μ_y? Compare the variance of $\tilde{\mu}_y$ to the variance of $\bar{y}$.

4. Assume that $y_i \sim (\mu_y, \sigma_y^2)$, that y_i is i.i.d., and that the sample size, n, is even. Let

$$\hat{\mu}_y = \frac{1}{2n} y_1 + \frac{3}{2n} y_2 + \frac{1}{2n} y_3 + \frac{3}{2n} y_4 + \cdots + \frac{1}{2n} y_{n-1} + \frac{3}{2n} y_n$$

Is $\hat{\mu}_y$ an unbiased estimator for μ_y? Compare the variance of $\hat{\mu}_y$ to the variance of $\bar{y}$.