Unit 11
Comparing Variances & Comparing Means

Including a Review of EDA

Review of Variance

(p. 11.1)

- Data set: {3, 4, 5, 8}
- Mean
  - $n = 4$
  - $\sum x = 3 + 4 + 5 + 8 = 20$
  - xbar = 20 / 4 = 5
- Variance and standard deviation
  - SS $=(x_i - \text{xbar})^2 = (3-5)^2 + (4-5)^2 + (5-5)^2 + (8-5)^2 = 14$
  - Variance (s^2) = SS / ($n-1$) = 14 / (4 − 1) ≐ 4.667
  - Standard deviation (s) $= \sqrt{s^2} = \sqrt{4.667} \doteq 2.2$
- Interpretation of standard deviation
  - Best when used to compare 2 groups
  - For normal distributions: 95% of data within $\mu \pm 2\sigma$
  - For all distributions: ≥75% within $\mu \pm 2\sigma$ (Chebychev)

Graphs

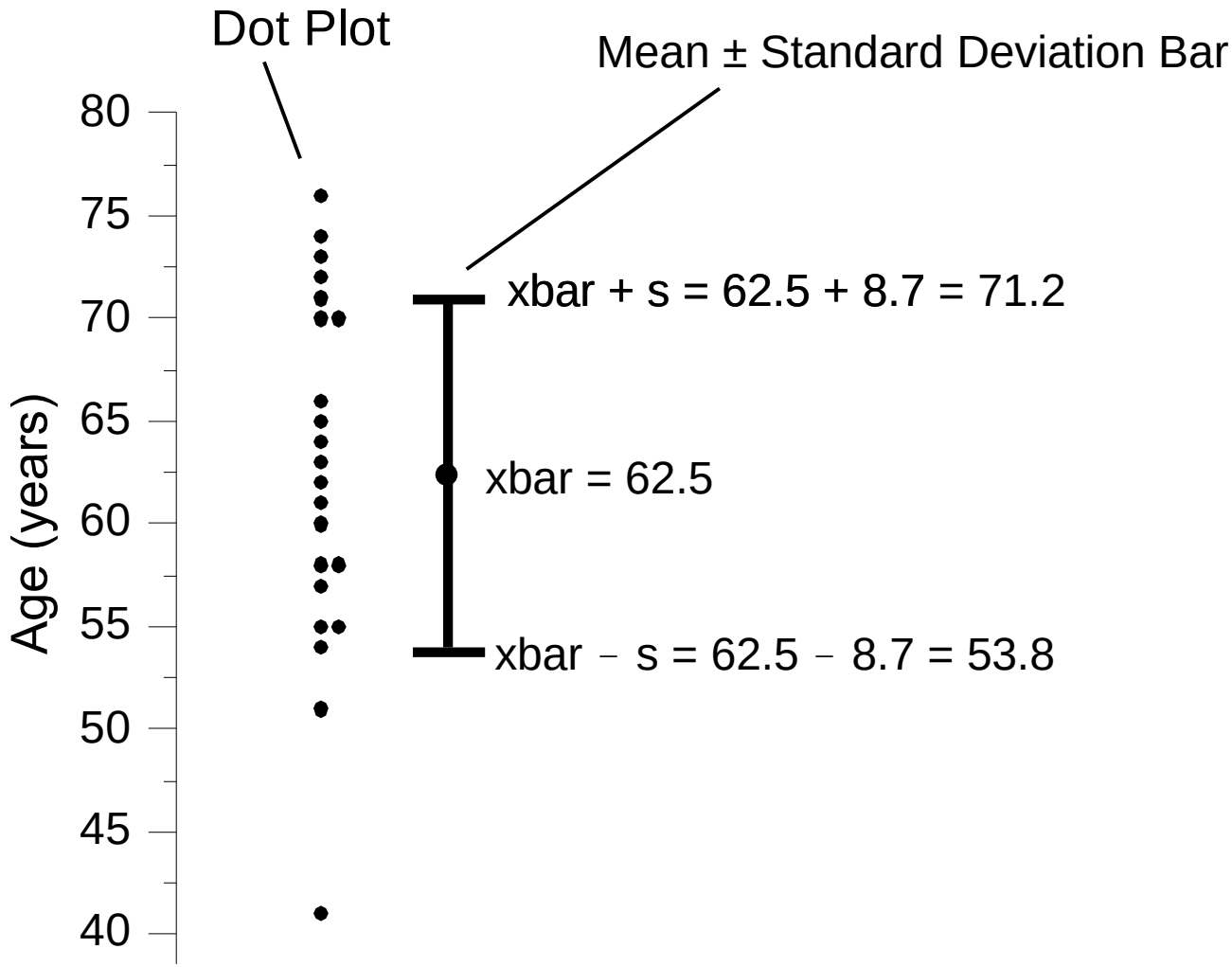
- Illustrative data: `agebycen.sav`, group 1 (age)
  - n = 22
  - xbar ≐ 62.5
  - s ≐ 8.7
- Stem-and-leaf ("stemplot")
  - draw stem first
  - Use between 4 and 12 bins

```
|4|1
|4|
|5|14
|5|55788
|6|01234
|6|56
|7|001234
|7|6
(x10)
```

This stemplot uses split (double) stem values to draw out the shape of the distribution.

Dot Plot and Mean±SD Plot

(p. 11.2)



5-Point Summary and Boxplot

p. 11.2 (cont.)

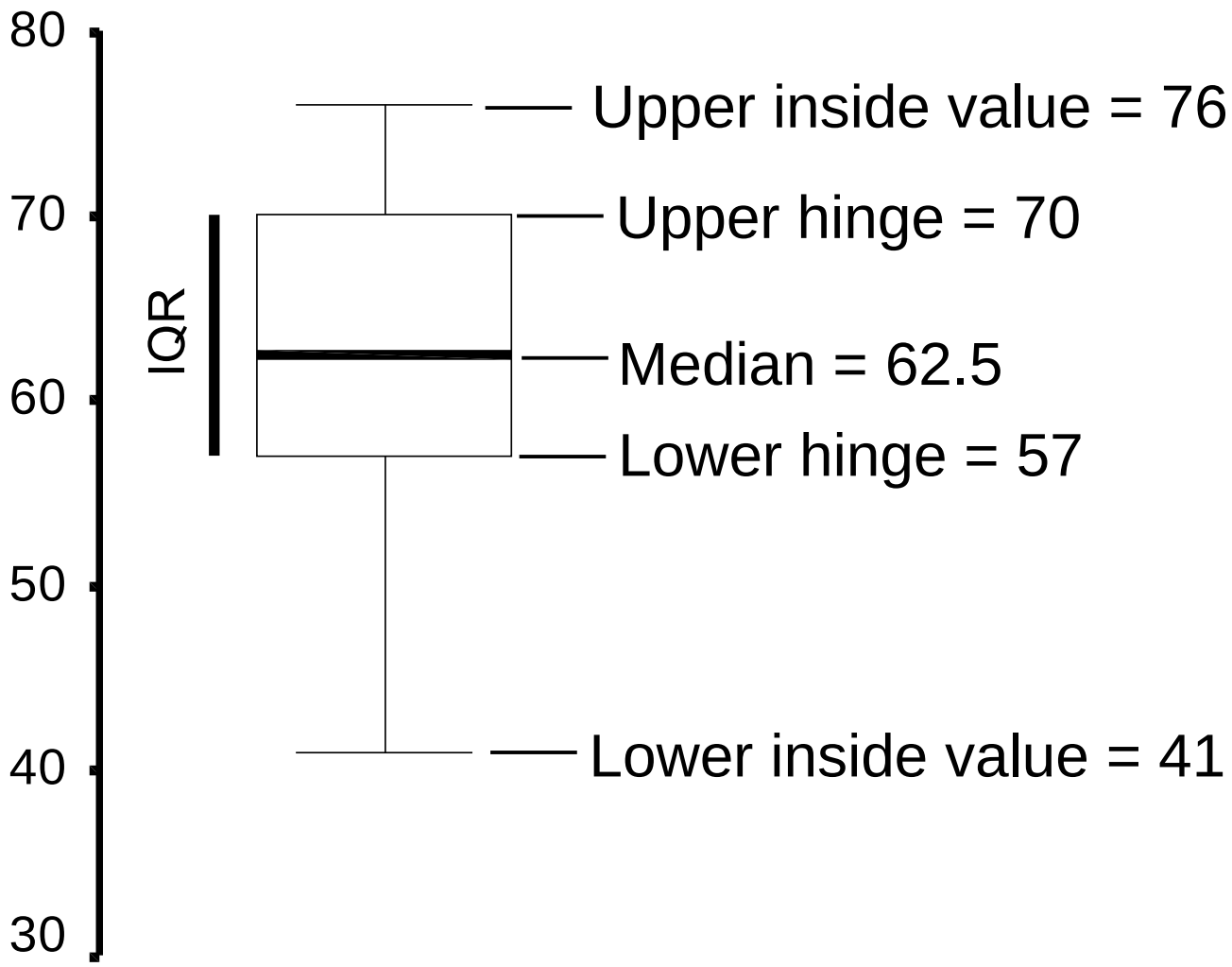
- Data (`agebycen.sav, group 1`)

```
41  51  54  55  55  57  58  58  60  61  62  63  64  65  66  70  70  71  72  73  74  76
Q0                  Q1                  Q2                      Q3                      Q4
```

- 5-point summary
  - Q0 = 41
  - Q1 = 57
  - Q2 = 62.5
  - Q3 = 70
  - Q4 = 76
- IQR = Q3−Q1= 70−57=13
- Check for upper outside values
  - $\text{Fence}_{\text{upper}}$ = Q3 + (1.5)(IQR) = 70 + (1.5)(13) = 89.5
  - no upper outside values
  - upper inside value = 76
- Check for lower outside values
  - $\text{Fence}_{\text{lower}}$ = Q1 − (1.5)(IQR) = 57 −(1.5)(13) = 37.5
  - no lower outside values
  - lower inside value = 41

Boxplot

p. 11.2 (cont.)



Confidence Interval for σ^2

p. 11.3

- We won't cover this technique but *note*:
- σ^2 = population variance
  - parameter
  - unknown
  - what we want to know
- s^2 = sample variance
  - estimate
  - calculated
  - What we have
- σ^2 and s^2 are related but are not the same!

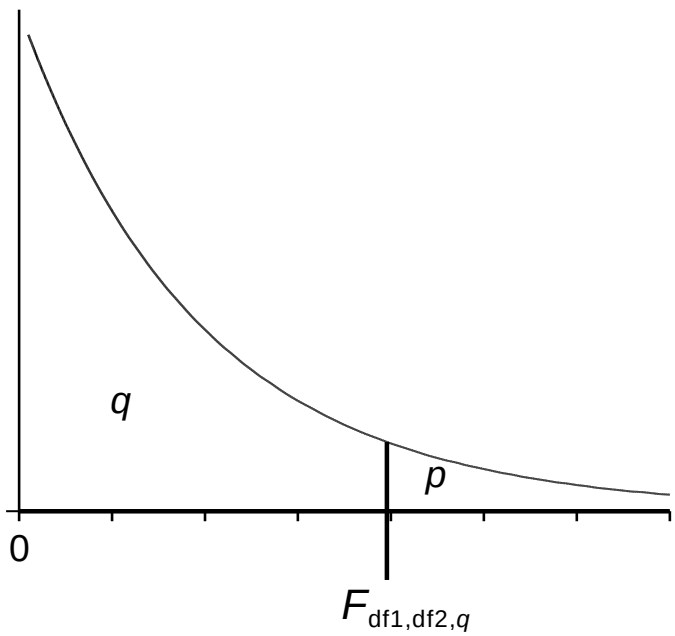
F ratio test
p. 11.4

- Illustrative data: independent samples
  - $s^2_1 = 4.667$, $n_1 = 4$
  - $s^2_2 = 4.333$, $n_2 = 4$
  - Do samples come from populations with equal variance?
- Test procedure (F ratio test)
  - H_0: $\sigma^2_1 = \sigma^2_2$
  - Flexible significance testing requires no preset α
  - $F_{stat} = s^2_1 / s^2_2$ or s^2_2 / s^2_1 , whichever larger
  - Convert F_{stat} to p value (next page)
    - Reject H_0 when p is "small"
    - $p < .1 \rightarrow$ some evidence against H_0
    - $p < .05 \rightarrow$ stronger evidence against H_0
    - $p < .01 \rightarrow$ still stronger evidence against H_0

F distributions

p. 11.5

- Used to determine probability of observing F_{stat} or greater
- df_1 = numerator degrees of freedom
- df_2 = denominator degrees of freedom
- Notation $F_{df1,\ df2,\ q}$ = F percentile with df_1, df_2, and cumulative probability of *q*



F table provides 95th percentile landmarks

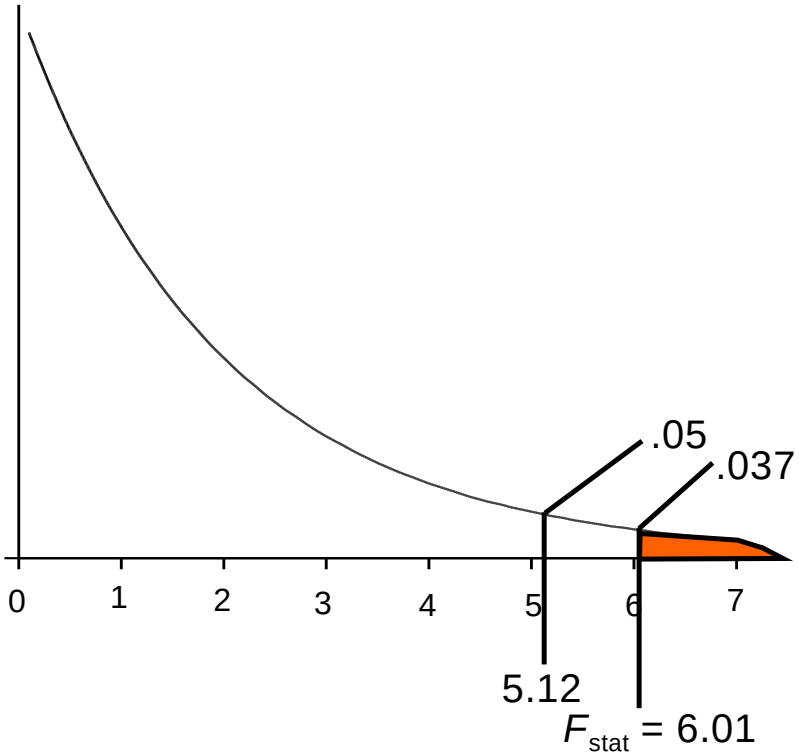
(p. 11.5)

		Numerator Degrees of Freedom										
	*	1	2	3	4	5	6	7	8	9	10	*
	1	161	199	216	225	230	234	237	239	241	242	1
	2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	2
D	3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	3
e	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	4
n	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	5
o												
m	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	6
i	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	7
n	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	8
a	9	*5.12*	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	9
t	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	10

$F_{1,9,.95} = 5.12$

Suppose $F_{stat} = 6.01$ with $df_1 = 1$ and $df_2 = 9$

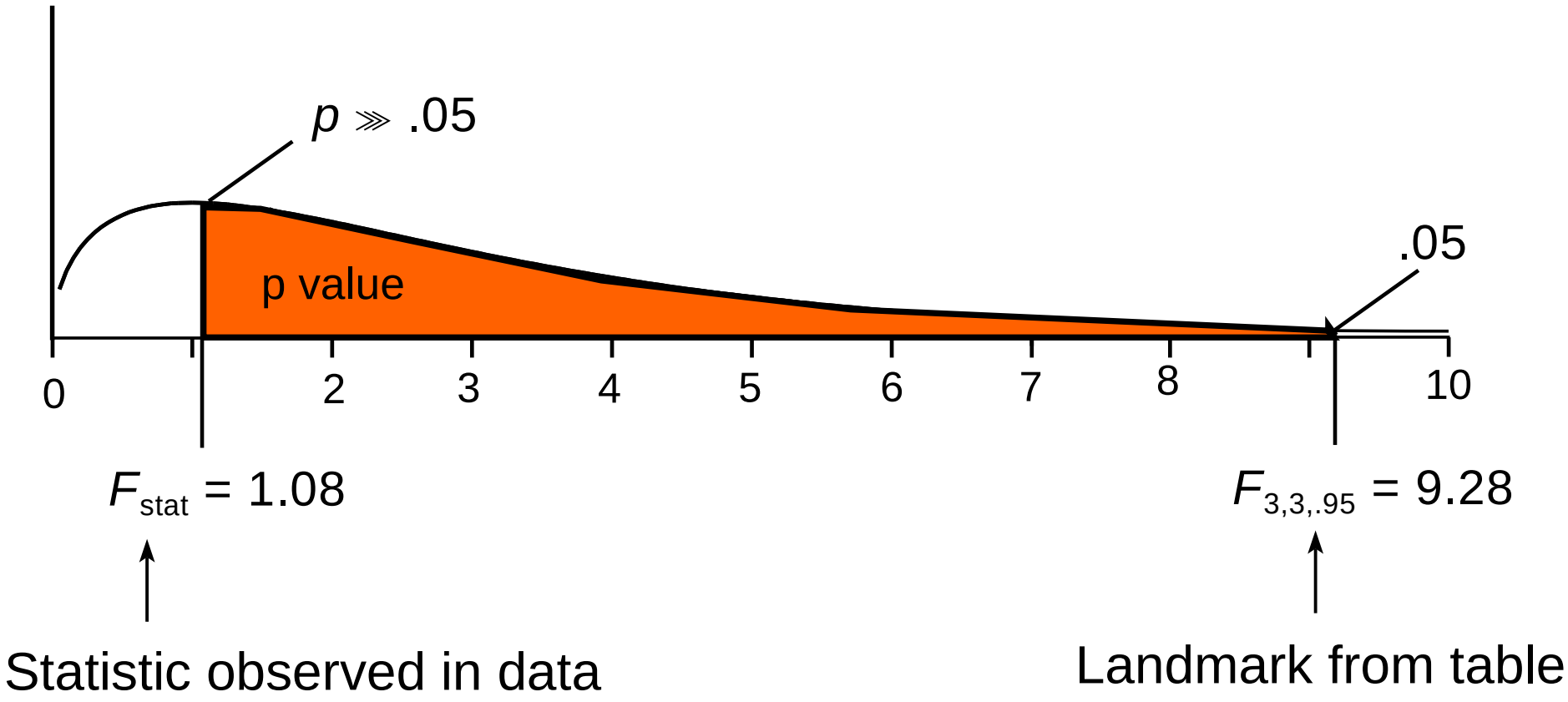
(p. 11.5)



- *p* value = shaded area in right tail
  - Method 1: use landmark from F table: $p < .05$
  - Method 2: use computer (e.g., StaTable): $p = .037$

Suppose $F_{stat} = 1.08$ with $df_1 = 3$ and $df_2 = 3$

(p. 11.6)



Pooling Variances

(p. 11.7)

- If $\sigma^2_1 \approx \sigma^2_2$ (equal variance, homoscedasticity) → combine ("pool") sample variances to estimate σ^2
- If $\sigma^2_1 \neq \sigma^2_2$ (unequal variance, heteroscadasticty) → do *not* pool sample variances
- How do you determine if $\sigma^2_1 \approx \sigma^2_2$?
  - Compare s_1 to s_2
  - EDA (e.g., side-by-side boxplots)
  - *F* ratio test

Example of Homoscedasticity

(not in Reader)

```
Group 1        Group 2
          0|9|
          0|8|
          0|7|0
          0|6|0
          0|5|0
          |4|0
          |3|0
          |2|
          |1|
         (x10)
```

$s_1 = 15.8$

$s_2 = 15.8$

Example of Heteroscedasticity

(not in Reader)

```
Group 1        Group 2
          0|9|0
          0|8|
          0|7|0
          0|6|
          0|5|0
          |4|
          |3|0
          |2|
          |1|0
         (x10)
```

$s_1 = 15.8$

$s_2 = 31.6$

Compare Means (equal variance assumed)

(p. 11.7)

- s^2_p = formula 7
  - But what is it?
  - ANS: weighted average of variances from the two samples
- $se_{\text{mean dif}}$ = formula 8
  - But what is it?
  - ANS: measure of precision of estimated mean dif
- t_{stat} = formula 9
  - But what question does it answer?
  - ANS: Can we say with confidence that populations means differ?

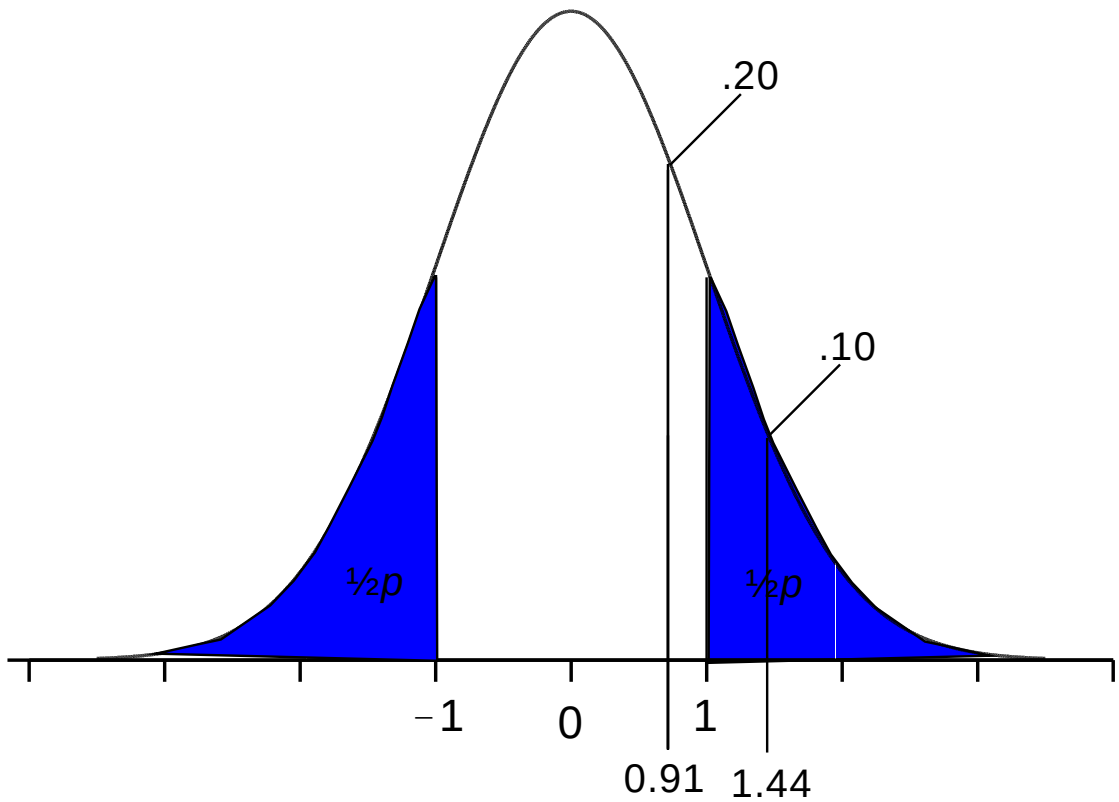
Comparing Means (equal variance assumed)

p. 11.7

- Simple illustrative example
  - $xbar_1 = 5$, $s_1^2 = 4.667$, $n_1 = 4$
  - $xbar_2 = 6.5$, $s_2^2 = 4.333$, $n_2 = 4$
- $s^2_p = [(3)(4.667) + (3)(4.333)] / 6 = 4.5$
- $se_{mean\ dif} = sqrt[4.5\times(1/4 + 1/4)] = 1.5$
- H_0: $\mu_1 = \mu_2$
  - $t_{stat} = (5 - 6.5) / 1.5 = -1.00$
  - df = (4 – 1) + (4 – 1) = 6
  - p value = .356 (see next slide)
  - Conclude: Difference is not significant.

Converting t_{stat} to p value

- $t_{stat} = -1.00$ with 6 df
- By hand (t table)
  - landmarks are $t_{6,.90} = 1.44$ and $t_{6,.80} = 0.91$
  - thus, $.10 < \frac{1}{2}p < .20$ and $.20 < p < .40$
- By computer: $p = .356$



Comparing Means (equal variance *not* assumed)

p. 11.8

- Do not pool variances
- Behrens-Fisher unequal variance t test statistic (formula 11)
- Standard Error (formula 12)
- if you do not have a computer that does the test for you use lesser of df_1 or df_2
- Illustrative example done on board