Final Review

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Hypothesis Testing

We make decisions every day in our lives.

Should I believe A or should I believe B (not A)?

Two Competing Hypotheses. A and B.

Null Hypothesis ($\mathbf{H_0}$): No difference, no association, or no effect.

Alternative Hypothesis ($\mathbf{H_1}$): Investigators belief.

The Alternative Hypothesis is always set up to be what you want to build up evidence to prove.

7.1 Introduction to Hypothesis Testing

The hypothesis testing process consists of 5 Steps.

Step 1: Set up the hypotheses and determine the level of significance.

Step 2: Select the appropriate test statistic.

Step 3: Set-up the decision rule.

Step 4: Compute the test statistic.

Step 5: Conclusion.

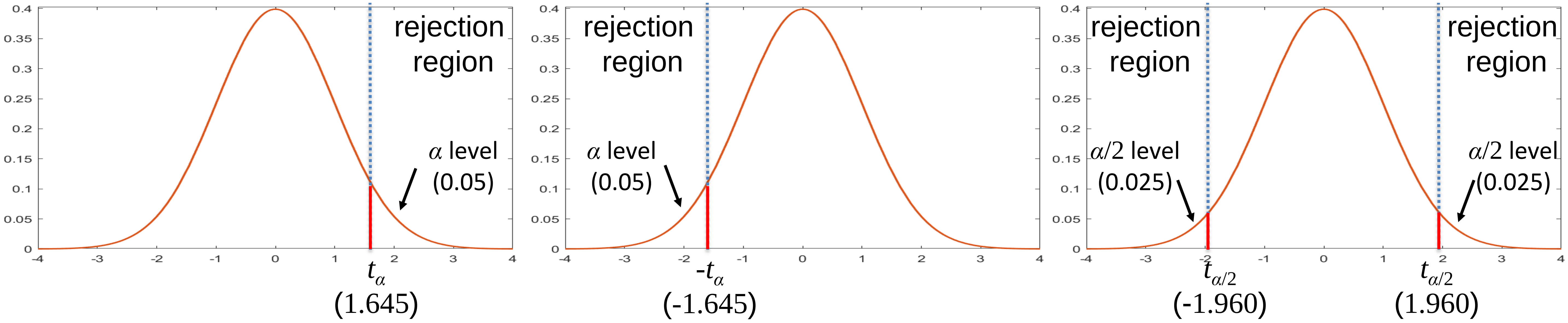
7.5 Tests with Two Independent Samples, Continuous Outcome

The hypothesis testing process consists of 5 Steps.

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{S_P\sqrt{1/n_1 + 1/n_2}}$$

Step 3: Set-up the decision rule.

$H_0: \mu_1=\mu_2$ vs. $H_1: \mu_1>\mu_2$ $\quad$ $H_0: \mu_1=\mu_2$ vs. $H_1: \mu_1<\mu_2$ $\quad$ $H_0: \mu_1=\mu_2$ vs. $H_1: \mu_1\neq\mu_2$



Reject H_0 if $t \geq t_{\alpha,df}$ $\quad$ Reject H_0 if $t \leq t_{\alpha,df}$ $\quad$ Reject H_0 $\leq t_{\alpha/2,df}$ or $t \geq t_{\alpha/2,df}$

Table 2 in book

7.5 Tests with Two Independent Samples, Continuous Outcome

	Sample Size	Mean	Standard Deviation
New drug	15	195.9	28.7
Placebo	15	227.4	30.3

Example: Is the mean cholesterol of new drug < mean of placebo?

Step 1: Null and Alternative Hypotheses.

H_0: $\mu_1 \geq \mu_2$ vs. H_1: $\mu_1 < \mu_2$

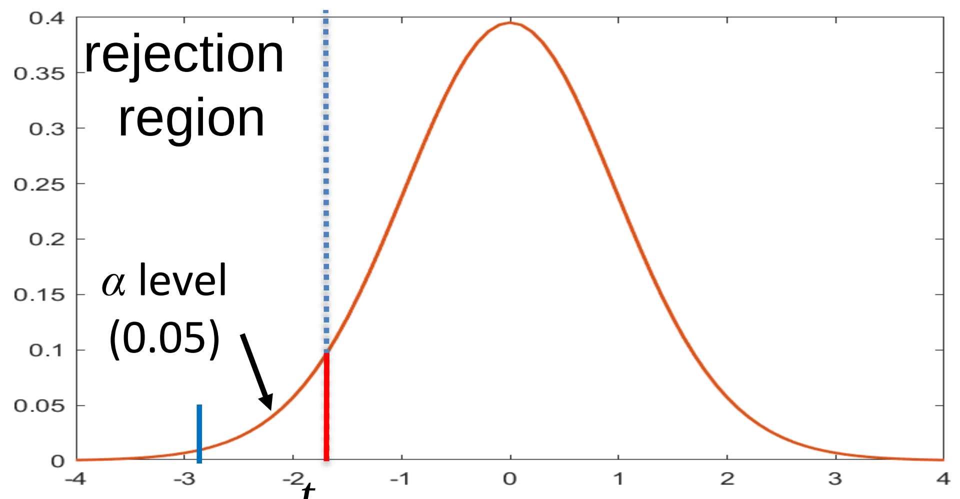
Step 2: Test Statistic.

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{S_P\sqrt{1/n_1 + 1/n_2}} \qquad df=n_1+n_2-2$$

Step 3: Decision Rule. $\alpha=0.05$, $df=15+15-2=28$

Reject H_0 if $t \leq -1.701$.

$df=28$



$t_{\alpha,df/2}$
(-1.701)

Step 4: Compute test statistic.

$$t = (195.9 - 227.4)\Big/\left(29.5\sqrt{1/15 + 1/15}\right) = -2.92$$

$$\bar{X}_i = \frac{1}{n}\sum X \qquad i=1,2$$

$$S_P = \sqrt{\frac{(15-1)(28.7)^2 + (15-1)(30.3)^2}{15+15-2}} = 29.5$$

Step 5: Conclusion

Because $-2.92 \leq -1.701$, reject and conclude mean of drug less than placebo.

7.6 Tests with Matched Samples, Continuous Outcome

The hypothesis testing process consists of 5 Steps.

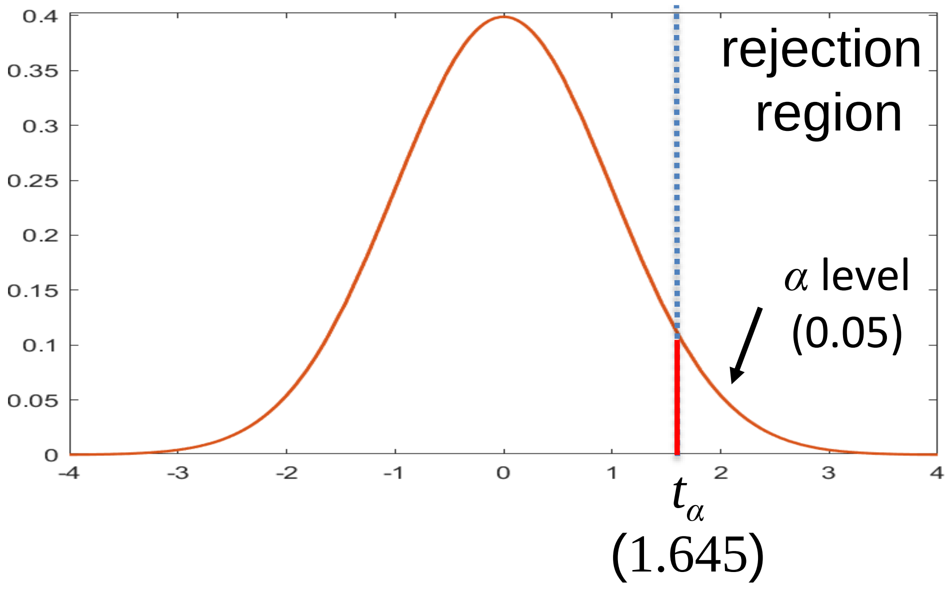
$$t = \frac{\bar{X}_d}{s_d / \sqrt{n}}$$

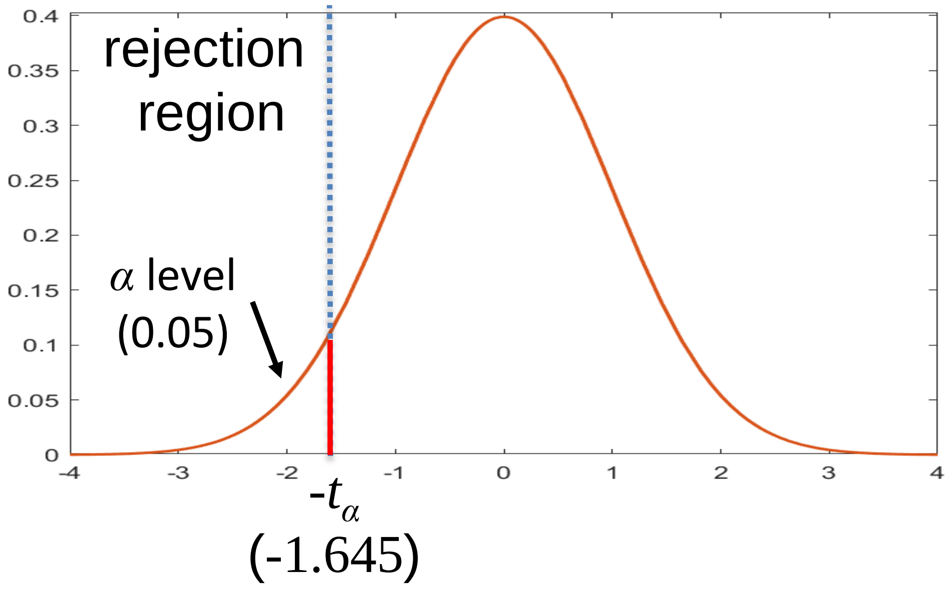
Step 3: Set-up the decision rule.

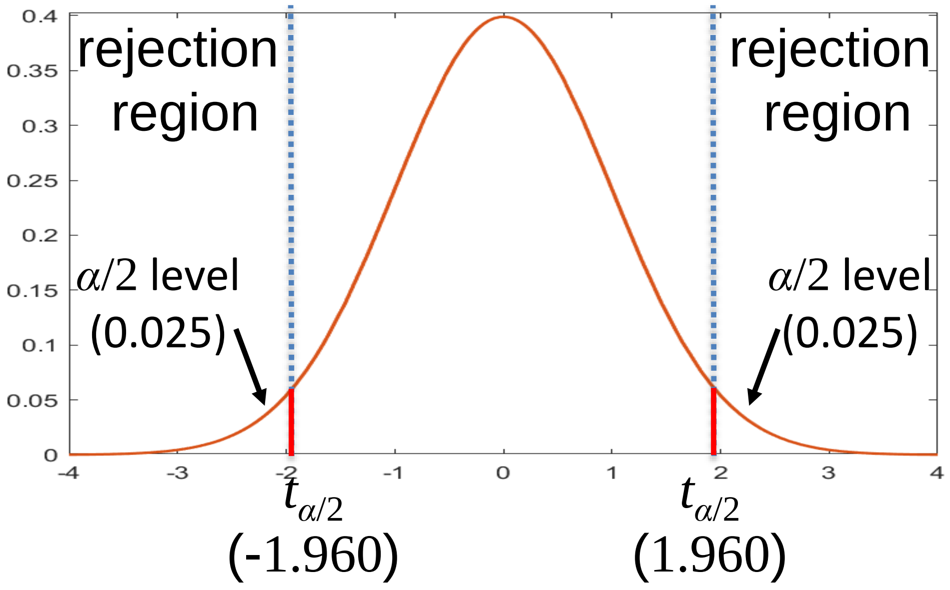
$H_0: \mu_d = 0$ *vs.* $H_1: \mu_d > 0$

$H_0: \mu_d = 0$ *vs.* $H_1: \mu_d < 0$

$H_0: \mu_d = 0$ *vs.* $H_1: \mu_d \neq 0$







Reject H_0 if $t \geq t_{\alpha,df}$

Table 2 in book

Reject H_0 if $t \leq t_{\alpha,df}$

Reject H_0 $t \leq t_{\alpha/2,df}$ or $t \geq t_{\alpha/2,df}$

7.6 Tests with Matched Samples, Continuous Outcome

Example: Is there a difference in mean of new drug from baseline?

Step 1: Null and Alternative Hypotheses.

$H_0: \mu_d = 0$ *vs*. $H_1: \mu_d \neq 0$

Step 2: Test Statistic.

$$t = \frac{\bar{X}_d}{s_d / \sqrt{n}} \qquad df = n\text{-}1$$

Step 3: Decision Rule. $\alpha = 0.05$, $df = 15\text{-}1 = 14$

Reject H_0 if $t \leq -2.145$ or $t \geq 2.145$.

Step 4: Compute test statistic.

$$t = -5.3 \big/ (12.8 / \sqrt{15}) = -1.60$$

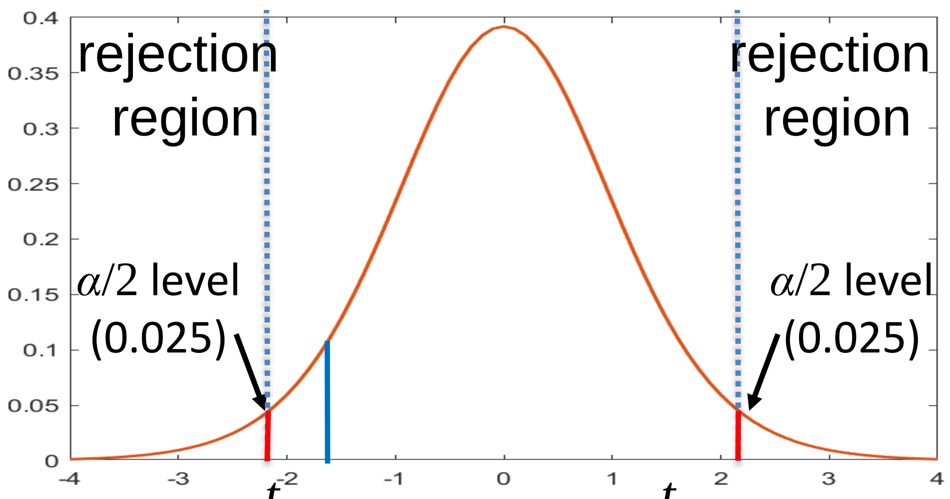
$$\bar{X}_d = \frac{1}{n} \sum d$$

$$s_d^2 = \frac{1}{n-1} \left[\sum d^2 - \frac{1}{n} \left(\sum d \right)^2 \right]$$

Step 5: Conclusion

Because $-2.145 \leq -1.60$, do not reject H_0 and conclude no reduction.

$df=14$



$t_{\alpha,df/2}$ (-2.145) $\qquad$ $t_{\alpha,df/2}$ (2.145)

Number	Baseline	6 Weeks	Difference
1	215	205	10
2	190	156	34
3	230	190	40
4	220	180	40
5	214	201	13
6	240	227	13
7	210	197	13
8	193	173	20
9	210	204	6
10	230	217	13
11	180	142	38
12	260	262	-2
13	210	207	3
14	190	184	6
15	200	193	7

$\bar{X}_d = -5.30$

7.7 Tests with Two Independent Samples, Dichotomous Outcome

We often have two populations that we are studying.

We may be interested in knowing if the proportion p_1 of population 1
is different (while accounting for random statistical variation)
from the proportion p_1 of population 2.

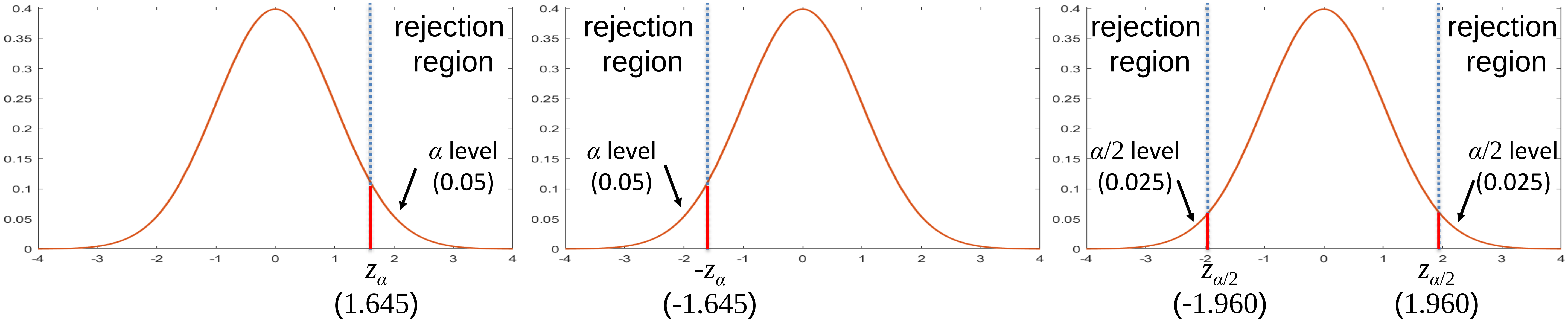
When we have independent random sample from each population
and the sample sizes are large.

7.7 Tests with Two Independent Samples, Dichotomous Outcome

The hypothesis testing process consists of 5 Steps.

Step 3: Set-up the decision rule. Assume n "Large."

$H_0: p_1=p_2$ vs. $H_1: p_1>p_2$ $\quad$ $H_0: p_1=p_2$ vs. $H_1: p_1<p_2$ $\quad$ $H_0: p_1=p_2$ vs. $H_1: p_1\neq p_2$



Reject H_0 if $z \geq z_\alpha$ $\quad$ Reject H_0 if $z \leq z_\alpha$ $\quad$ Reject H_0 $z \leq z_{\alpha/2}$ or $z \geq z_{\alpha/2}$

Table 1 in book

7.8 Tests with More than Two Independent Samples, Continuous Outcome (ANOVA)

The hypothesis testing process consists of 5 Steps.

Step 1: Set up the hypotheses and determine the level of significance α.

H_0: $\mu_1 = \mu_2 \ldots = \mu_k$ *vs*. H_1: at least two μ's different

reject for "large" disparities or $F = MSB/MSE$.

We will assume the means are equal and calculate two different variances.

If the means are truly equal, the two different variances will be the same.

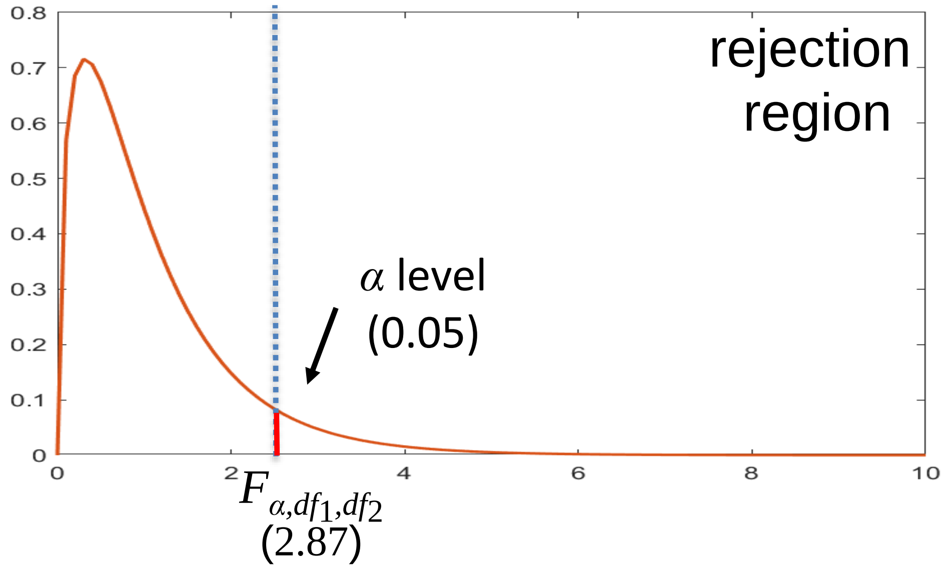
If the means are noy equal, the two different variances will be different.

7.8 Tests with More than Two Independent Samples, Continuous Outcome (ANOVA)

The hypothesis testing process consists of 5 Steps.

Step 3: Set-up the decision rule.

H_0: $\mu_1 = \mu_2 \ldots = \mu_k$ vs. H_1: at least two different



$df_1 = k$-1

$df_2 = N$-k

$$MSB = \frac{\sum n_j (\bar{X}_j - \bar{X})^2}{k-1}$$

$$MSE = \frac{\sum\sum n_j (X - \bar{X}_j)^2}{N-k}$$

$$F = \frac{MSB}{MSE}$$

Reject H_0 if $F \geq F_{\alpha,df1,df2}$.

Table 4 in book

See Chapter 07b worksheet for details

7.8 Tests with More than Two Independent Samples, Continuous Outcome (ANOVA)

Example: Find the value of $F_{0.05,3,16}$.

α $\quad df_1=n_1$-1 $\quad df_2=n_2$-1

The (critical) value of F that has an area of 0.05 larger than it when we have $df_1=3$ (numerator) and $df_2=16$ (denominator) degrees of freedom is 3.24.

$P(F_{df_1,df_2} > F) = 0.05$,
e.g., $P(F_{3,20} > 3.10) = 0.05$

	df_1													
df_2	1	2	3	4	5	6	7	8	9	10	20	30	40	50
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	248.0	250.1	251.1	251.8
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.45	19.46	19.47	19.48
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.66	8.62	8.59	8.58
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.80	5.75	5.72	5.70
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.56	4.50	4.46	4.44
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.87	3.81	3.77	3.75
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.44	3.38	3.34	3.32
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.15	3.08	3.04	3.02
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	2.94	2.86	2.83	2.80
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.77	2.70	2.66	2.64
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.65	2.57	2.53	2.51
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.54	2.47	2.43	2.40
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.46	2.38	2.34	2.31
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.39	2.31	2.27	2.24
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.33	2.25	2.20	2.18
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.28	2.19	2.15	2.12
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.23	2.15	2.10	2.08
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.19	2.11	2.06	2.04
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.16	2.07	2.03	2.00
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.12	2.04	1.99	1.97

This is the value we use for a 95% HT when $\alpha=0.05$, $n_1=6$, and $n_2=11$.

The book only has $\alpha=0.05$, but would have another page for each α value.

7.7 Tests with Two Independent Samples, Dichotomous Outcome

The hypothesis test on risk difference

H_0: $p_1 = p_2$ *vs*. H_1: $p_1 \neq p_2$

H_0: $RD = 0$ *vs*. H_1: $RD \neq 0$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Is equivalent to the two hypothesis tests

Risk Ratio RR

H_0: $RR = 1$ *vs*. H_1: $RR \neq 1$

$$RR = \frac{\hat{p}_1}{\hat{p}_2}$$

and

Odds Ratio OR

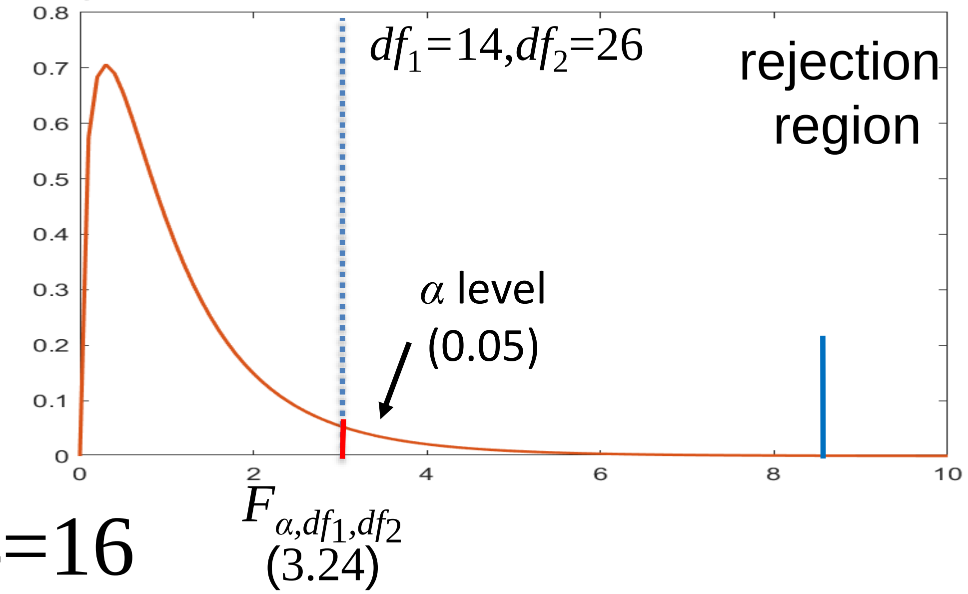
H_0: $OR = 1$ *vs*. H_1: $OR \neq 1$

$$OR = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)}$$

7.8 Tests with More than Two Independent Samples, Continuous Outcome (ANOVA)

Low-Calorie	Low-Fat	Low-Carbohydrate	Control
8	2	3	2
9	4	5	2
6	3	4	-1
7	5	2	0
3	1	3	3
6.6	3.0	3.4	1.2

Example: Statistical difference in weight loss among 4 diets?



Step 1: Null and Alternative Hypotheses.

H_0: $\mu_1=\mu_2=\mu_3=\mu_4$ *vs.* H_1: at least two different

Step 2: Test Statistic.

$F = MSB / MSE$ $df_1=k\text{-}1$ $df_2=N\text{-}k$

Step 3: Decision Rule. $\alpha=0.05$, $df_1=4\text{-}1=3$, $df_2=20\text{-}4=16$

Reject H_0 if $F \geq 3.24$.

$n_1=n_2=n_3=n_4=5$

Step 4: Compute test statistic.

F to be calculated

$$MSB = \frac{\sum n_j(\bar{X}_j - \bar{X})^2}{k-1} = 25.3$$

$$MSE = \frac{\sum\sum n_j(X - \bar{X}_j)^2}{N-k} = 3.0$$

$F = 25.3 / 3.0 = 8.43$

Step 5: Conclusion

Because $8.43 > 3.24$, reject H_0 and conclude diets mean weight loss different.

7.10 Summary

TABLE 7-50 Summary of Key Formulas for Tests of Hypothesis

Outcome Variable, Number of Groups: Null Hypothesis	Test Statistic*
Continuous outcome, two independent samples: $H_0: \mu_1 = \mu_2$	$z = \dfrac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{1/n_1 + 1/n_2}}$
Continuous outcome, two matched samples: $H_0: \mu_d = 0$	$z = \dfrac{\bar{X}_d - \mu_d}{s_d / \sqrt{n}}$
Continuous outcome, more than two independent samples: $H_0: \mu_1 = \mu_2 = \dots = \mu_k$	$F = \dfrac{\Sigma n_j(\bar{X}_j - \bar{X})^2 / (k-1)}{\Sigma\Sigma(X - \bar{X}_j)^2 / (N-K)}$
Dichotomous outcome, one sample: $H_0: p = p_0$	$z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$
Dichotomous outcome, two independent samples: $H_0: p_1 = p_2$, RD = 0, RR = 1, OR = 1	$z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$
Categorical or ordinal outcome, one sample: $H_0: p_1 = p_{10}, p_2 = p_{20}, \dots, p_k = p_{k0}$	$\chi^2 = \Sigma \dfrac{(O-E)^2}{E}, df = k-1$
Categorical or ordinal outcome, two or more independent samples: H_0 : Outcome and groups are independent	$\chi^2 = \Sigma \dfrac{(O-E)^2}{E}, df = (r-1)(c-1)$

Associations

We often are interested in the association between variables.

We often say **correlation**, with little thought to an actual definition.

We often say trend or **linear** relationship without defining how determine this relationship.

We define y to be the response or **dependent** (on x) **variable** and x to be the explanatory or **independent variable**. i.e. y depends on x (or several x's).

Associations

We often are interested in the association between variables.

We often say **correlation**, with little thought to an actual definition.

We often say trend or **linear** relationship without defining how determine this relationship.

We define y to be the response or **dependent** (on x) **variable** and x to be the explanatory or **independent variable**. i.e. y depends on x (or several x's).

9.3 Introduction to Correlation and Regression Analysis-Correlation

5 Sums $\sum X \quad \sum Y$

$\sum X^2 \quad \sum Y^2$

$\sum XY$

Correlations r are between -1 and 1, $-1 \le r \le 1$.

$$r = \frac{\operatorname{cov}(x,y)}{\sqrt{s_x^2 s_y^2}}$$

Variance of X

$$s_x^2 = \frac{1}{n-1}\sum (X-\bar{X})^2 = \frac{1}{n-1}\left[\sum X^2 - \frac{1}{n}\left(\sum X\right)^2\right]$$

Variance of Y

$$s_y^2 = \frac{1}{n-1}\sum (Y-\bar{Y})^2 = \frac{1}{n-1}\left[\sum Y^2 - \frac{1}{n}\left(\sum Y\right)^2\right]$$

CoVariance of X&Y

$$\operatorname{cov}(x,y) = \frac{1}{n-1}\sum (Y-\bar{Y})(X-\bar{X}) = \frac{1}{n-1}\left[\sum XY - \frac{1}{n}\left(\sum Y\right)\left(\sum X\right)\right]$$

Not in book

9.3 Introduction to Correlation and Regression Analysis-Correlation

We are going to calculate the correlation in column format with sums.

n	X	X^2	Y	Y^2	XY
1	34.7	1204.1	1895	3591025.0	65756.5
2	36.0	1296.0	2030	4120900.0	73080.0
3	29.3	858.5	1440	2073600.0	42192.0
4	40.1	1608.0	2835	8037225.0	113683.5
5	35.7	1274.5	3090	9548100.0	110313.0
6	42.4	1797.8	3827	14645929.0	162264.8
7	40.3	1624.1	3260	10627600.0	131378.0
8	37.3	1391.3	2690	7236100.0	100337.0
9	40.9	1672.8	3285	10791225.0	134356.5
10	38.3	1466.9	2920	8526400.0	111836.0
11	38.5	1482.3	3430	11764900.0	132055.0
12	41.4	1714.0	3657	13373649.0	151399.8
13	39.7	1576.1	3685	13579225.0	146294.5
14	39.7	1576.1	3345	11189025.0	132796.5
15	41.1	1689.2	3260	10627600.0	133986.0
16	38.0	1444.0	2680	7182400.0	101840.0
17	38.7	1497.7	2005	4020025.0	77593.5
	652.1	25173.2	49334.0	150934928.0	1921162.6

5 Sums

$$\sum X = 652.1 \qquad \sum X^2 = 25173.2$$

$$\sum Y = 49334.0 \qquad \sum Y^2 = 150934928.0$$

$$\sum XY = 1921162.6$$

$$\text{cov}(x,y) = 1798.0$$

$$s_x^2 = 9.9638$$

$$s_y^2 = 485478.8$$

$$r = \frac{1798.0}{\sqrt{(10.0)(485478.8)}}$$

$$r = 0.82$$

9.3 Introduction to Correlation and Regression Analysis-Regression

We can estimate the y-intercept and slope from what we have already computed for the correlation.

$s_x^2 = 9.9638$

$s_y^2 = 485478.8$

$r = 0.82$

The slope is estimated as $b_1 = r\dfrac{s_y}{s_x}$ and $b_0 = \bar{Y} - b_1\bar{X}$.

Point slope formula

Line goes through $(\bar{X},\bar{Y})$. Note b_1 has same sign as r.

And hence we have determined our regression line.

$\hat{y} = b_0 + b_1 x$

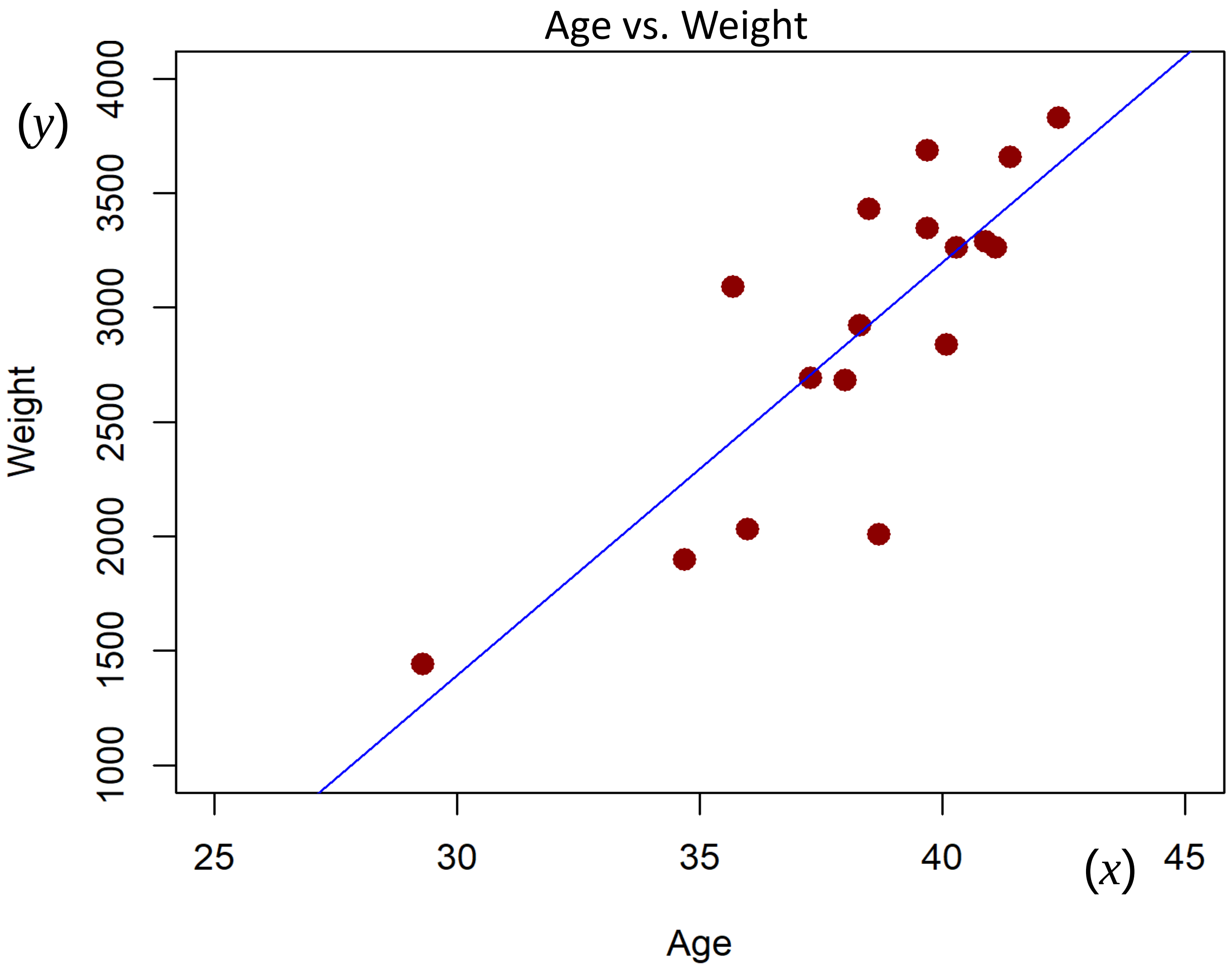
9.3 Introduction to Correlation and Regression Analysis-Regression

Example: Continuing the small study … to investigate the association between gestational age and birth weight.

$$s_x^2 = 9.9638$$

$$s_y^2 = 485478.8$$

$$r = 0.82$$



$$b_1 = r\frac{s_y}{s_x}$$

$$b_1 = 0.82\frac{696.8}{3.2}$$

$$b_1 = 180.5$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$b_0 = 2902 - (180.5)(38.4)$$

$$b_0 = -4029.2$$

$$\hat{Y} = -4029.2 + 180.5x$$

9.4 Multiple Linear Regression Analysis

Example: SBP and BMI, Age, Male Sex, and TFH.

A multiple regression analysis is run and coefficients estimated.

$$SBP = 68.15 + 0.58BMI + 0.65AGE + 0.94MLS + 6.44TFH$$

Independent Variable	Regression Coefficient	t	p-value
Intercept	b_0=68.15	t_0=26.33	0.0001=p_0
BMI	b_1= 0.58	t_1=10.30	0.0001=p_1
Age	b_2= 0.65	t_2=20.22	0.0001=p_2
Male sex	b_3= 0.94	t_3= 1.58	0.1133=p_3
Treatment for hypertension	b_4= 6.44	t_4= 9.74	0.0001=p_4

You will often see this type of output.

The t statistic is for H_0: $\beta_j=0$, H_1: $\beta_j \neq 0$.

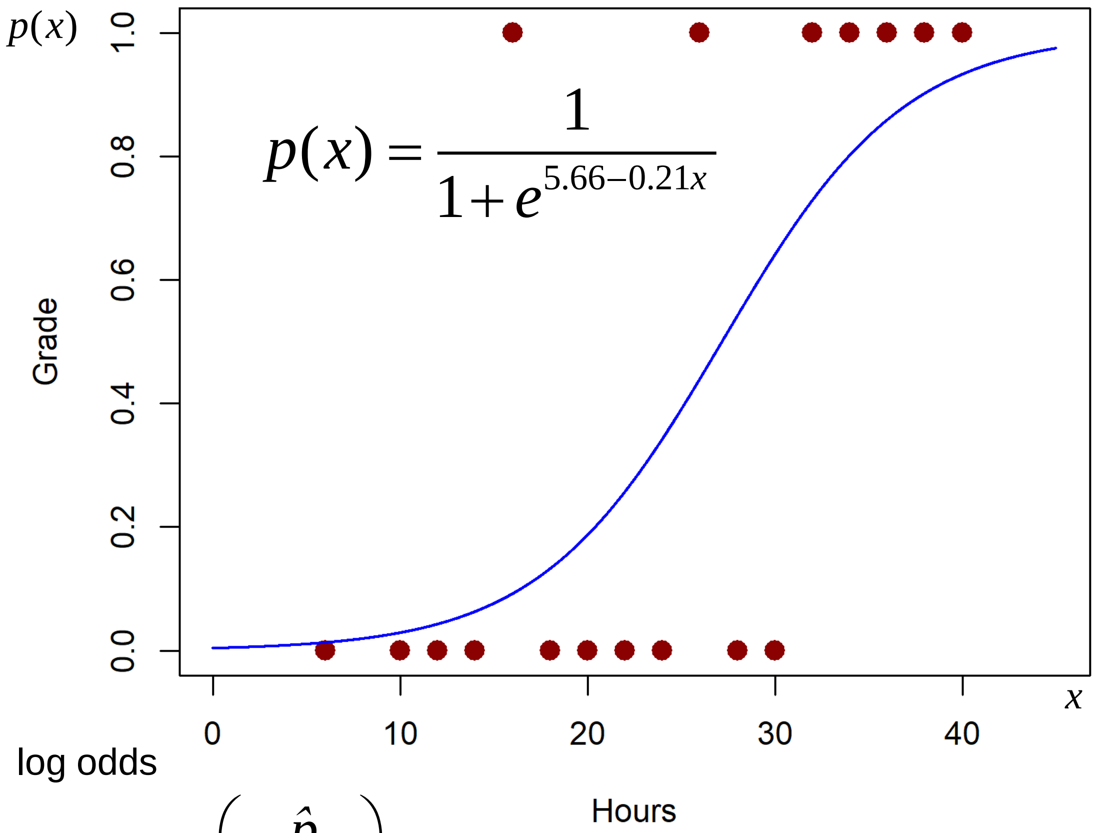
The *p-value* is the probability of getting this coefficient estimate or larger in abs if it were truly 0.

$$t_j = \frac{b_j - 0}{\sqrt{\mathrm{var}(b_j)}}$$

$df=n-p-1$

9.5 Multiple Logistic Regression Analysis

Using ***R***,



$$p(x)=\frac{1}{1+e^{5.66-0.21x}}$$

log odds

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right)=-5.66+0.21x$$

output

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.65549    2.54651  -2.221   0.0264
xx           0.20778    0.09302   2.234   0.0255
```

Hours (x)	A (y)
6	0
8	0
10	0
12	0
14	0
16	1
18	0
20	0
22	0
24	0
26	1
28	0
30	0
32	1
34	1
36	1
38	1
40	1

```
# grade data
xx <- c(6, 8,10,12,14,16,18,20,22,24,26,28,30,32,34,36,38,40)
yy <- c(0, 0, 0, 0, 0, 1, 0, 0, 0, 0 ,1, 0, 0, 1, 1, 1, 1, 1)

#scatter plot plot(x = xx,y = yy,xlab = “Hours",ylab = “Grade",
xlim = c(0,45),ylim = c(0,1),col = "darkred",
  cex = 1.5, main = "Hours vs. Grade", pch = 16)

logistic_model <- glm(yy~xx, family=binomial(link="logit"))
summary(logistic_model)
b0    <- logistic_model$coefficients[1]
b1    <- logistic_model$coefficients[2]
phat <- round(1/(1+exp(-b0-b1*xx)), digits = 4)
O     <- round(phat/(1-phat)             , digits = 4)
df    <- data.frame(xx,yy,phat,O)
df

#scatter plot with curve
xhat <- (1:4500)/100
yhat <- 1/(1+exp(-b0-b1*xhat))
plot(x = xx,y = yy,xlab = “Hours",ylab = “Grade",
  xlim = c(0,45),ylim = c(0,1),col = "darkred",
  cex = 1.5, main = "Hours vs. Grade", pch = 16)
points(xhat,yhat,cex = .1,col = "blue")
```

9.5 Multiple Logistic Regression Analysis

Once we have $\hat{\beta}_0$ and $\hat{\beta}_1$, insert them back into

$$\hat{p}_i = \frac{1}{1+e^{-\hat{\beta}_0 - \hat{\beta}_1 x_i}}$$ for estimated probabilities

and also for odds $\hat{o}_i = \frac{\hat{p}_i}{1-\hat{p}_i} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}$

OR for a difference in x

and for odds ratio $\hat{OR} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_b} / e^{\hat{\beta}_0 + \hat{\beta}_1 x_a} = e^{\hat{\beta}_1 \Delta}$, $\Delta = x_b - x_a$.

$$\hat{\beta}_0 = -5.66$$
$$\hat{\beta}_1 = \quad 0.21$$

$$\hat{OR} = e^{(0.21)(2)} = 1.5220$$

OR for a difference of x=2

Study 2 more hours and *OR* increases by 1.5.

Hours (x)	A (y)	$\hat{p}$	$\hat{o}$
6	0	0.0120	0.0122
8	0	0.0181	0.0184
10	0	0.0272	0.0279
12	0	0.0406	0.0423
14	0	0.0603	0.0641
16	1	0.0886	0.0972
18	0	0.1284	0.1473
20	0	0.1824	0.2232
22	0	0.2527	0.3381
24	0	0.3388	0.5124
26	1	0.4371	0.7764
28	0	0.5405	1.1764
30	0	0.6406	1.7824
32	1	0.7298	2.7008
34	1	0.8036	4.0923
36	1	0.8611	6.2008
38	1	0.9038	9.3957
40	1	0.9344	14.2365

9.6 Summary

Correlation

$$\text{cov}(x,y)=\frac{1}{n-1}\left[\sum XY-\frac{1}{n}\left(\sum Y\right)\left(\sum X\right)\right]$$

$$s_x^2=\frac{1}{n-1}\left[\sum X^2-\frac{1}{n}\left(\sum X\right)^2\right]$$

$$s_y^2=\frac{1}{n-1}\left[\sum Y^2-\frac{1}{n}\left(\sum Y\right)^2\right]$$

$$r=\frac{\text{cov}(x,y)}{\sqrt{s_x^2 s_y^2}}$$

Linear Regression

$$b_1=r\frac{s_y}{s_x}$$

$$\hat{y}=b_0+b_1x$$

$$b_0=\bar{Y}-b_1\bar{X}$$

Logistic Regression

$$\hat{p}=\frac{1}{1+e^{-b_0-b_1x_1-\ldots-b_px_p}}$$

logistic probability

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right)=b_0+b_1x_1+\ldots+b_px_p$$

log odds

$$\hat{O}R=e^{\hat{\beta}_1\Delta_1+\ldots+\hat{\beta}_p\Delta_p}$$

odds ratio for difference Δ_j in x_j

10.1 Introduction to Nonparametric Testing – Sign Test

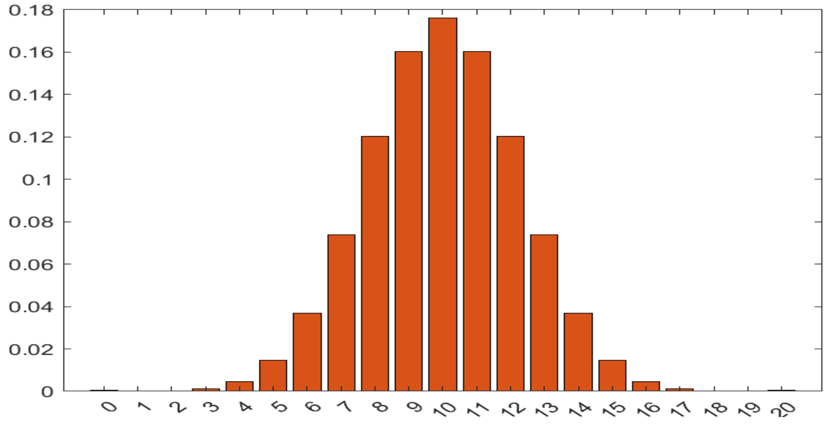
Example: Mark is training for 10K. n=20 daily runs.

Step 1: $H_0: MD=4$ *vs*. $H_1: MD>4$, $\alpha=0.05$

Step 2: Test Statistic.

x = the number of +'s.

Binomial Distribution, $n=20$, $p=0.5$



$X_{0.05}=15$ (or n-5=15)

x	P(X=x)	CumSum	CumSumR
0	0.000	0.000	1.000
1	0.000	0.000	1.000
2	0.000	0.000	1.000
3	0.001	0.001	1.000
4	0.005	0.006	0.999
5	0.015	0.021	0.994
6	0.037	0.058	0.979
7	0.074	0.132	0.942
8	0.120	0.252	0.868
9	0.160	0.412	0.748
10	0.176	0.588	0.588
11	0.160	0.748	0.412
12	0.120	0.868	0.252
13	0.074	0.942	0.132
14	0.037	0.979	0.058
15	0.015	0.994	0.021
16	0.005	0.999	0.006
17	0.001	1.000	0.001
18	0.000	1.000	0.000
19	0.000	1.000	0.000
20	0.000	1.000	0.000

Table 6

Two-Sided Test α	.10	.05	.02	.01
One-Sided Test α	.05	.025	.01	.005
n				
1				
2				
3				
4				
5	0			
6	0	0		
7	0	0	0	
8	1	0	0	0
9	1	1	0	0
10	1	1	0	0
11	2	1	1	0
12	2	2	1	1
13	3	2	1	1
14	3	2	2	1
15	3	3	2	2
16	4	3	2	2
17	4	4	3	2
18	5	4	3	3
19	5	4	4	3
20	5	5	4	3
21	6	5	4	4
22	6	5	5	4
23	7	6	5	4
24	7	6	5	5
25	7	7	6	5

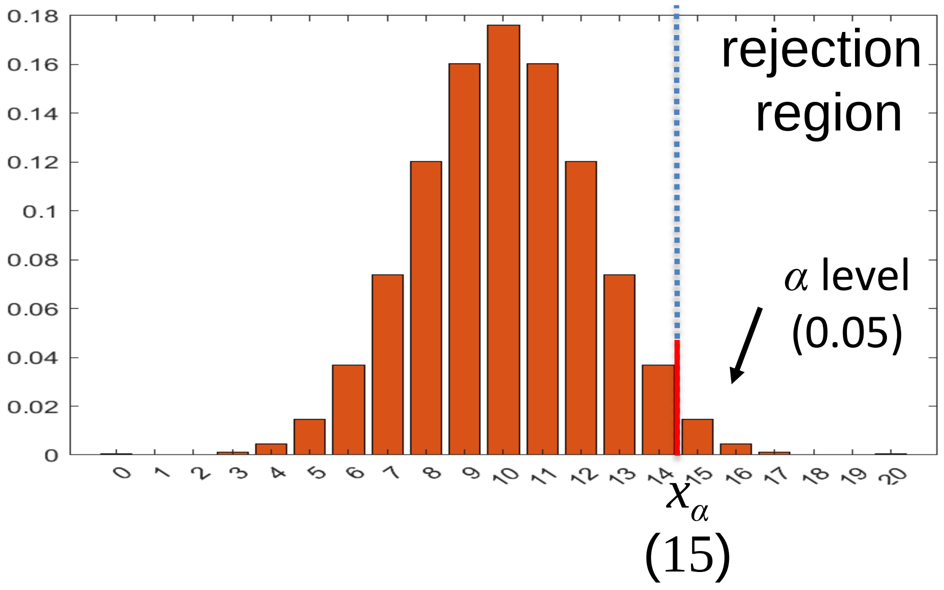
data
5
3
5
3
4
4
6
6
6
4
6
5
5
5
4
5
5
5
5
6

10.1 Introduction to Nonparametric Testing – Sign Test

The hypothesis testing process consists of 5 Steps.

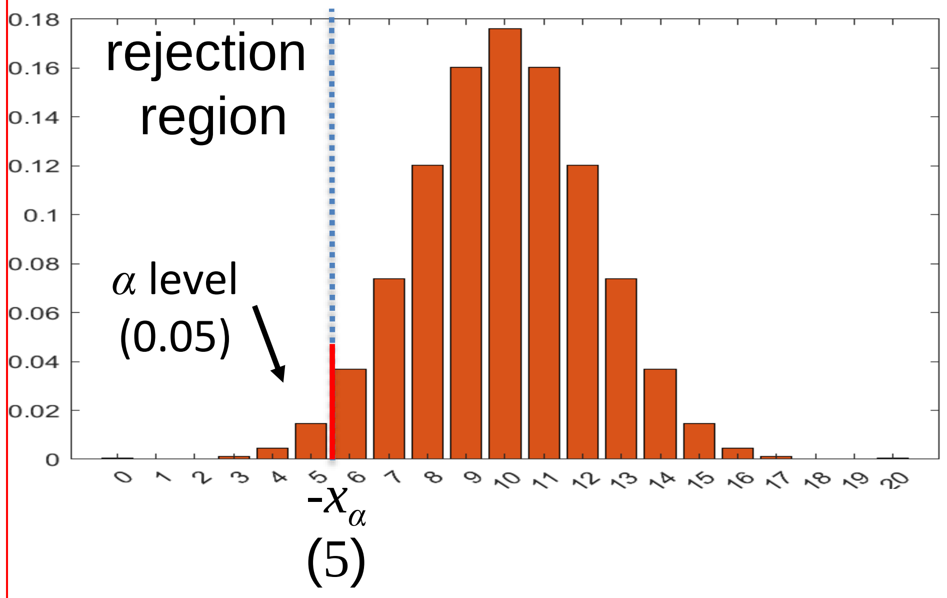
Step 3: Set-up the decision rule.

$\mathrm{H}_0: MD=4$ *vs*. $\mathrm{H}_1: MD>4$



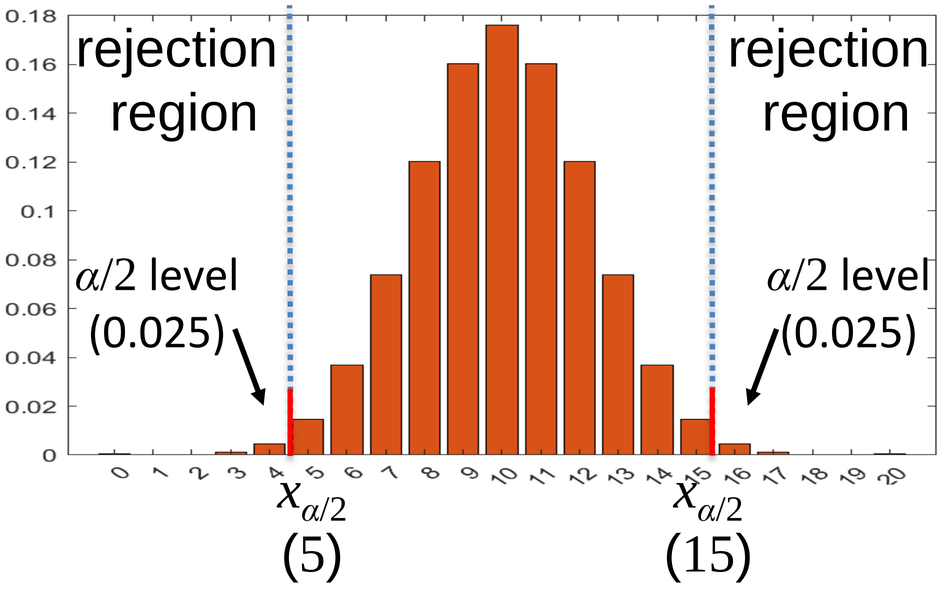
Reject H_0 if $P(X \geq x_\alpha) \leq \alpha$

$\mathrm{H}_0: MD=4$ *vs*. $\mathrm{H}_1: MD<4$



Reject H_0 if $P(X \leq x_\alpha) \leq \alpha$

$\mathrm{H}_0: MD=4$ *vs*. $\mathrm{H}_1: MD \neq 4$



Reject H_0 $P(x_\alpha \leq X) \leq \dfrac{\alpha}{2}$

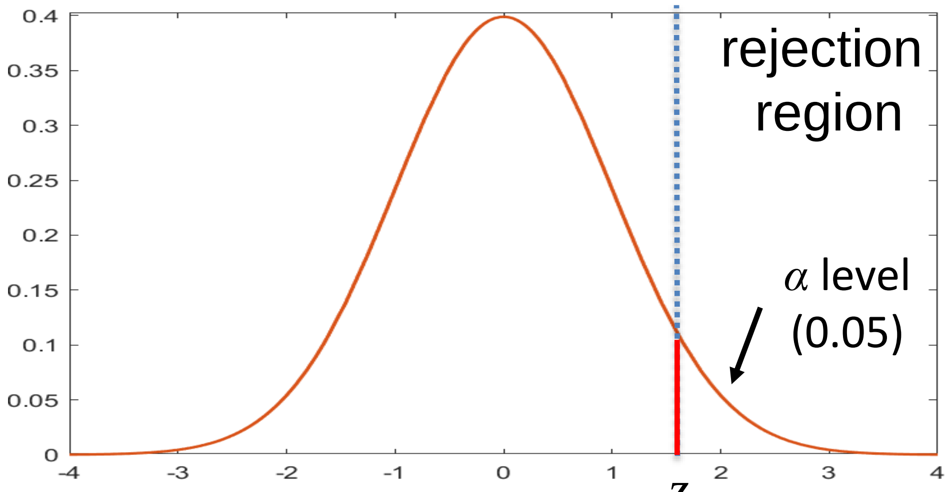
7.1 Introduction to Hypothesis Testing

RECALL

The hypothesis testing process consists of 5 Steps.

Step 3: Set-up the decision rule.

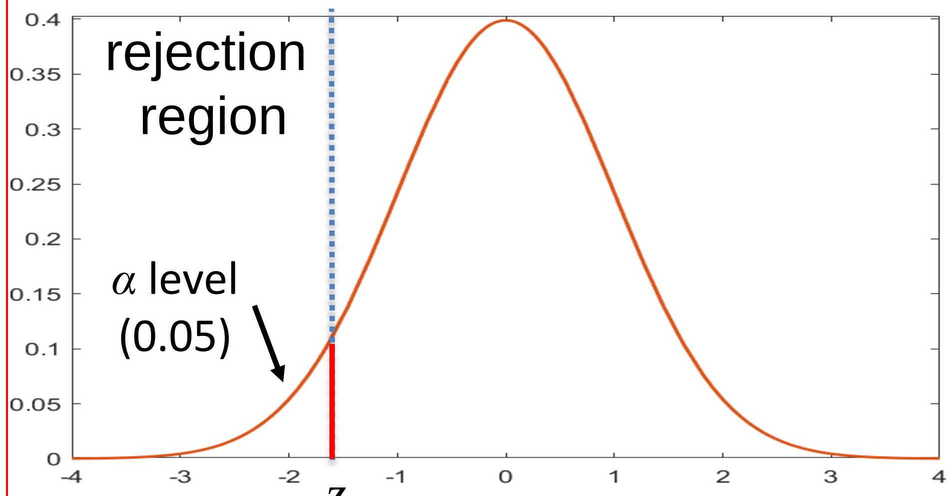
$\mathrm{H}_0: \mu = \mu_0$ *vs.* $\mathrm{H}_1: \mu > \mu_0$



z_α
(1.645)

Reject H_0 if $z \geq z_\alpha$

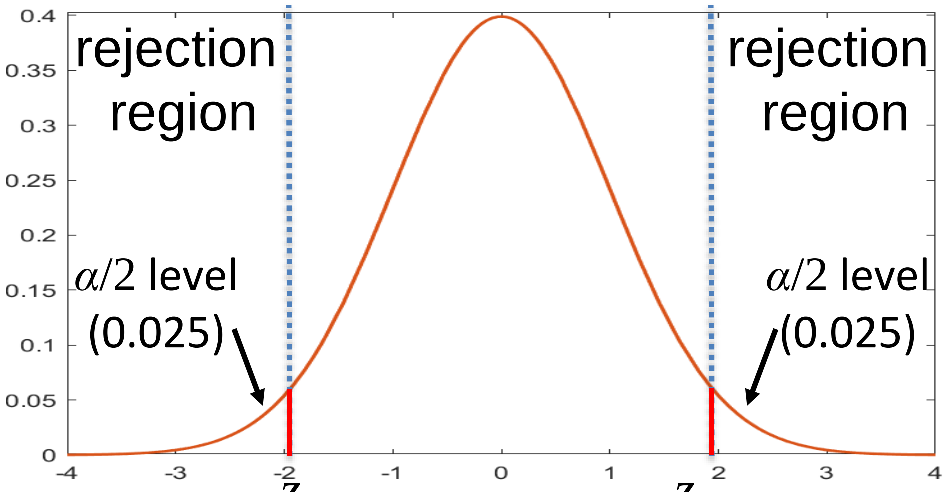
$\mathrm{H}_0: \mu = \mu_0$ *vs.* $\mathrm{H}_1: \mu < \mu_0$



$-z_\alpha$
(-1.645)

Reject H_0 if $z \leq z_\alpha$

$\mathrm{H}_0: \mu = \mu_0$ *vs.* $\mathrm{H}_1: \mu \neq \mu_0$



$-z_{\alpha/2}$
(-1.960)

$z_{\alpha/2}$
(1.960)

Reject H_0 $z \leq z_{\alpha/2}$ or $z \geq z_{\alpha/2}$

10.1 Introduction to Nonparametric Testing – Sign Test

If value < MD_0, –.
If value = MD_0, 0.
If value > MD_0, +.

data	sorted	sign
5	3	-1
3	3	-1
5	4	0
3	4	0
4	4	0
4	4	0
6	5	+1
6	5	+1
6	5	+1
4	5	+ 1
6	5	+1
5	5	+1
5	5	+1
5	5	+1
4	5	+1
5	6	+1
5	6	+1
5	6	+1
5	6	+1
6	6	+1

Step 4: Compute the test statistic.

$x = 14$

x = (the number of observations > MD_0=4)

Step 5: Because x=14 < x_α=15, do not reject H_0.

x	P(X=x)	CumSum	CumSumR
5	0.015	0.021	0.994
6	0.037	0.058	0.979
14	0.037	0.979	0.058
15	0.015	0.994	0.021

See also Table 6

Two-Sided Test α	.10	.05	.02	.01
One-Sided Test α	.05	.025	.01	.005
19	5	4	4	3
20	5	5	4	3
21	6	5	4	4

Table 6

Note:

If we used normal, we would reject H_0, t=4.07>$t_{0.05,19}$=2.093.

$t = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$ df=n-1 $\bar{X} = 4.8500$ $s = 0.9333$

10.2 Tests with Two Independent Samples – Mann-Whitney U Test

Example: Phase II clinical trial, n=10 children. Difference in episodes?

Group 1	Group 2
Placebo	NewDrug
7	3
5	6
6	4
4	2
12	10

$n_1 = 5 \quad n_2 = 5$

Step 1: Set up the hypotheses and determine α.

$$\mathrm{H}_0{:}MD_1 = MD_2 \text{ vs. } \mathrm{H}_1{:}MD_1 \neq MD_2, \quad \alpha = 0.05$$

Step 2: Select the appropriate test statistic.

Pool data and assign ranks. Test statistic based on ranks

Placebo	New Drug	Placebo	New Drug	Placebo	New Drug	Ranks Placebo	Ranks New Drug
7	3		1		1		1
5	6		2		2		2
6	4		3		3		3
4	2	4	4	4.5	4.5	4.5	4.5
12	1	5		6		6	
		6	6	7.5	7.5	7.5	7.5
		7		9		9	
		12		10		10	

$R_1 = 37$

$R_2 = 18$

10.2 Tests with Two Independent Samples – Mann-Whitney U Test

Step 2: Select the appropriate test statistic.

The test statistic is a single (decision) number summarizing information.

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = (5)(5) + \frac{5(5+1)}{2} - 37 = 3$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = (5)(5) + \frac{5(5+1)}{2} - 18 = 22$$

$$U = \min(U_1, U_2) = \min(3, 22) = 3$$

Reject H_0 for small U.

Rankings

Group 1	Group 2
	1
	2
	3
	4
	5
6	
7	
8	
9	
10	

$U = 0$

Group 1	Group 2
	1
2	
	3
4	
	5
6	
	7
8	
	9
10	

$U = 25$

complete separation

complete alternating

$$0 \le U \le n_1 n_2$$

10.2 Tests with Two Independent Samples – Mann-Whitney U Test

Step 3: Set-up the decision rule.

n_1=5, n_2=5

If we did Two Sided Test

Reject H_0 if $U \leq U_{0.05,n_1,n_2}$.

Step 4: Compute test statistic.

Already done, U=3.

Step 5: Conclusion.

Do not reject H_0 because

$U=3>U_{0.05,5,5}=2$. Interpret.

Two-Sided Test $\alpha = 0.05$

	n_1																			
n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2								0	0	0	0	1	1	1	1	1	2	2	2	2
3					0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4				0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5			0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6			1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7			1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8		0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9		0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10		0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11		0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12		1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13		1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14		1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15		1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16		1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17		2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18		2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19		2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20		2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

10.3 Tests with Matched Samples – Wilcoxon Signed Rank Test

An alternative for the Sign Test for matched samples median difference is the Wilcoxon Signed Rank test.

Step 1:

H_0: The median difference is zero (H_0: $\delta=0$)

H_1: The median difference is positive (H_1: $\delta>0$)

H_0: $\delta\leq0$ *vs*. H_1: $\delta > 0$
δ is population version of d.

We will calculate a test statistic W the smaller of W+ and W– .

$W+$ = sum of positive ranks

W_- = sum of negative ranks

$\longrightarrow$ $W=\min(W+, W_-)$

If the median difference of the matched pairs is zero, then the sum of the positive ranks should be the same as the sum of the negative ranks.

10.3 Tests with Matched Samples – Wilcoxon Signed Rank Test

An alternative for the Sign Test for matched samples median difference is the Wilcoxon Signed Rank test.

Step 1:

$H_0: \delta \leq 0$ *vs*. $H_1: \delta > 0$

δ is population version of d.

b	a	d	sorted	sign	rank	SgnRnk
85	75	10	-10	-1	3	-3
70	50	20	-5	-1	1	-1
40	50	-10	10	+1	3	3
65	40	25	10	+1	3	3
80	20	60	15	+1	5	5
75	65	10	20	+1	6	6
55	40	15	25	+1	7	7
20	25	-5	60	+1	8	8

$n=8$

Step 2: Select the test statistic.

$$W_+ = \text{sum of positive ranks} = 32$$

$$W_- = \text{sum of negative ranks} = 4$$

$$W = \min(W_1, W_2) = \min(4, 32) = 4$$

Reject H_0 for small W.

IF **Signed Ranks**

SgnRnk	SgnRnk	SgnRnk	SgnRnk
1	-4	-7	-8
2	-3	-5	-7
3	-2	-3	-6
4	-1	-1	-5
5	5	2	2
6	6	4	4
7	7	6	6
8	8	8	8
$W=0$	$W=10$	$W=16$	$W=26$

Possible Examples

10.3 Tests with Matched Samples – Wilcoxon Signed Rank Test

Step 3: Set-up the decision rule.

n=8, α=0.05

If we did One Sided Test

Reject H_0 if $W \leq W_{\alpha,n}$.

Two-Sided Test α	.10	.05	.02	.01
One-Sided Test α	.05	.025	.01	.005
n				
5	1			
6	2	1		
7	4	2	0	
8	6	4	2	0
9	8	6	3	2
10	11	8	5	3

Table 7

Step 4: Compute test statistic.

Already done, W=4.

Step 5: Conclusion.

Reject H_0 because

$W=4 \leq W_{0.05,8}=6$. Interpret.

10.4 Tests with More than Two Independent Samples – Kruskal-Wallis Test

The hypothesis testing process consists of 5 Steps.

Step 1: Set up the hypotheses and determine the level of significance α.

H_0: $MD_1 = MD_2 \ldots = MD_k$ *vs*. H_1: at least two MD's different
reject for “large" disparities H.

We will assume the medians are equal and see how different from equal.

7.8 Tests with More than Two Independent Samples, Continuous Outcome (ANOVA)

RECALL

The hypothesis testing process consists of 5 Steps.

Step 1: Set up the hypotheses and determine the level of significance α.

H_0: $\mu_1 = \mu_2 \ldots = \mu_k$ *vs*. H_1: at least two μ's different

reject for "large" disparities $F = MSB/MSE$.

We will assume the means are equal and calculate two different variances.

If the means are truly equal, the two different variances will be the same.

If the means are noy equal, the two different variances will be different.

10.4 Tests with More than Two Independent Samples – Kruskal-Wallis

Example: Statistical difference in albumin for 3 diets?

Step 1: Null and Alternative Hypotheses.

H_0: $MD_1=MD_2=MD_3$ *vs*. H_1: at least two different

$$H = \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(N+1)$$

Step 2: Test Statistic.

$$H = \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(N+1)$$

Step 3: Decision Rule. $\alpha=0.05$, $n_1=3$, $n_2=5$, $n_3=4$

Reject H_0 if $H \geq 5.656$.

Table 8

Three groups				
n_1	n_2	n_3	$\alpha = .05$	$\alpha = .01$
5	4	3	5.656	7.445

Sample size order doesn’t matter.

Step 4: Compute test statistic.

$$H = 7.52$$

Step 5: Conclusion

Reject H_0 because $7.52 > 5.656$, and conclude difference in median albumin.

10.5 Summary

Sign Test (one sample)

x = number of observations > MD_0

Mann-Whitney U Test

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

$$U = \min(U_1, U_2)$$

Sign Test (two sample)

x = number of observations > 0

Wilcoxon Signed Rank Test

$W = \min(W+, W-)$

$W+$ = sum of positive ranks

$W-$ = sum of negative ranks

Kruskal-Wallis Test

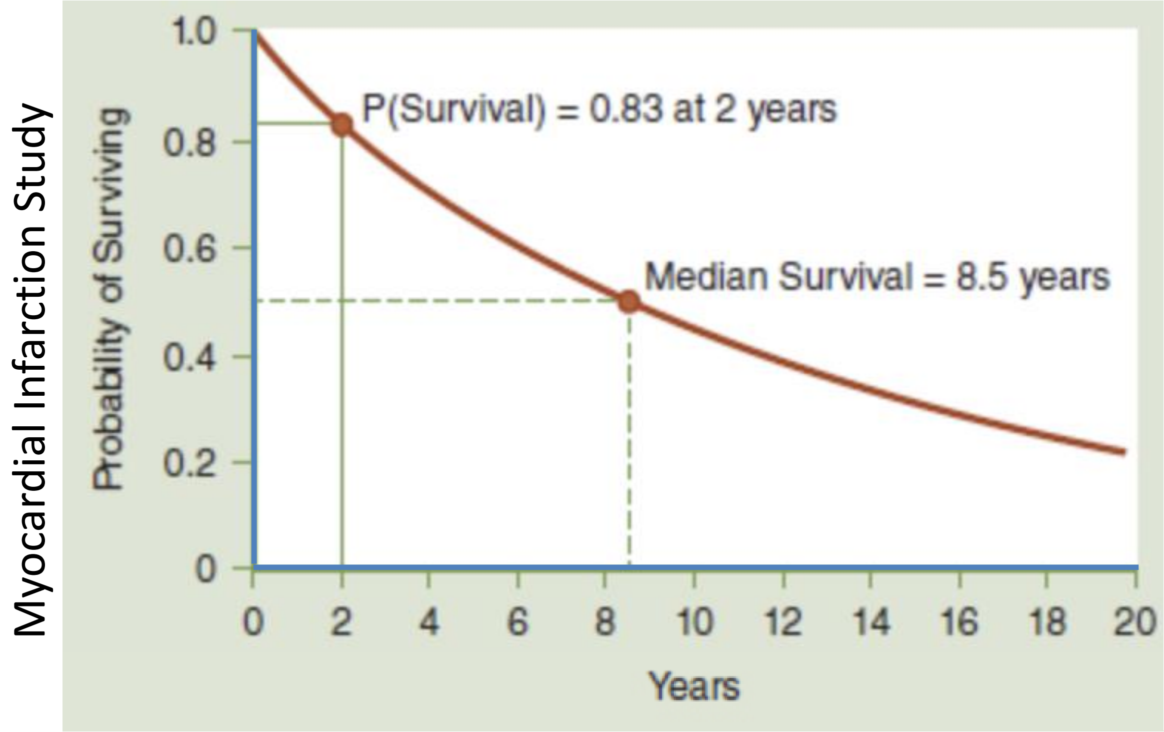
$$H = \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(N+1)$$

11.1 Introduction to Survival Data

Survival analysis measures two pieces of information

1) Whether the event occurred, 1=yes, 0=no
2) Last follow-up time, from enrollment.

The **survival function** is the probability a person survives past a time t.



t=0.0 : survival probability=1.00

t=2.0 : survival probability=0.83

t=8.5 : survival probability=0.50 (Median)

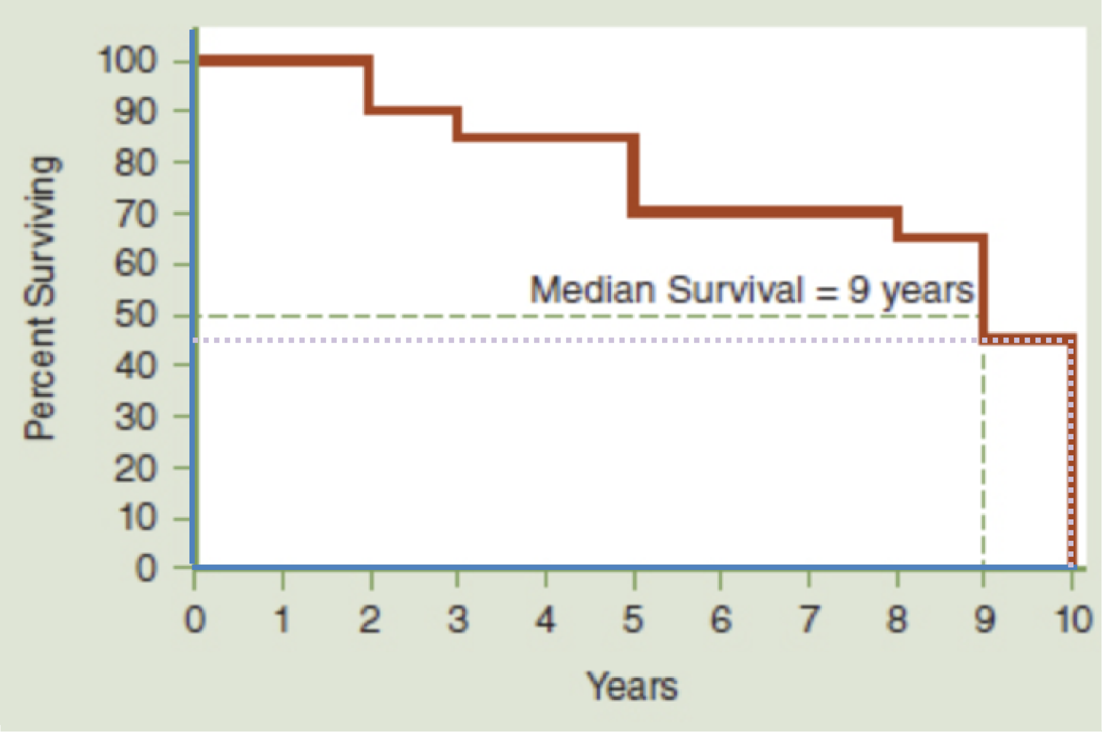
t=10.0: survival probability=0.47

11.2 Estimating the Survival Function

There are several parametric and nonparametric ways to estimate survival curves.

Let's examine nonparametric step survival curves.

Time on x axis and survival (percentage) at risk on y axis.



t=0.0 : survival probability=1.00

t=2.0 : survival probability=0.90

t=9.0 : survival probability=0.50 (Median)

t=10.0: survival probability=0.45

11.6 Summary

The **survival function** is the probability a person survives past a time t.

~~Actuarial Life Table~~

N_t = # event free during interval t
(Number at risk)

D_t = # who die in interval t

C_t = # censored in interval t

N_{t*} = avg. # at risk in interval t, $N_{t*}=N_t-C_t/2$

q_t = prop. die in interval t, $q_t=D_t/N_{t*}$

p_t = prop. survive in interval t, $p_t=1-q_t$

S_t = prop. survive past interval t

Can plot S_t vs. t.

~~Kaplan-Meier Life Table~~

$$S_{t+1} = S_t \frac{N_t - D_t}{N_t}$$

$$SE(S_t) = S_t \sqrt{\sum \frac{D_t}{N_t(N_t - D_t)}}$$

~~Chi-Square Test~~

$$\chi^2 = \sum_{j=1}^{2} \frac{\left(\sum_{t=1}^{T} O_{ij} - \sum_{t=1}^{T} E_{ij}\right)^2}{\sum_{t=1}^{T} E_{ij}} \qquad df = k-1$$

~~Cox Proportional Hazards Model~~

$$h(t) = h_0(t)\exp(b_1x_1 + b_2x_2 + \ldots + b_px_p)$$

Questions?

Bring pencil/eraser, calculator, caffeinated beverage.

Will hand out exam and formula sheet/tables.