Stat 135 lec 21 (no lec 20)

Remaining 2 quizzes pushed back 1 week

Quiz 3: Nov 8

Quiz 4: Nov 22

Last time

sec 11.2 Comparing the means of independent normal populations with the same variance.

$X_1,\dots,X_n \sim N(\mu_x,\sigma^2)$, $Y_1,\dots,Y_m \sim N(\mu_y,\sigma^2)$

σ^2 known:

$\bar{X}-\bar{Y} \sim N\left(\mu_x-\mu_y, \sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right)$

2 sample Z-test.

$$Z = \frac{\bar{X}-\bar{Y}-(\mu_x-\mu_y)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim N(0,1)$$

σ^2 unknown:

we use unbiased estimator of σ^2:

Pooled sample variance

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

(2 sample t-test)

$$t = \frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$$

if $n+m-2$ is ≥ 30 you can use 2 sample Z-test.

Today

0. Duality of HT and CI.
1. Comparing means of independent normal populations with different variances (Sec 11.2)
2. Power calculations (Sec 11.2)
3. Comparing paired Samples (Sec 11.3)

Post Midterm topics

- parametric test (Normal theory t-test) comparing the means of 2 normal populations (Sec 11.2, 11.3)

- non parametric test comparing the means of 2 continuous populations. (Sec 11.2, 11.3)

- Chi square tests testing whether 2 or more categorical variables have the same distribution.

- ANOVA (Normal theory F-test) comparing the means of 2 or more normal populations (Sec 12.2)

- Regression (Chap 14)

- Bayesian Statistics (parameter estimation, hypothesis testing)

ex (Problem 11.6.10) (Duality of HT and CI)

Verify that the two sample t-test at level α of $H_0: \mu_X - \mu_Y = 0$

$H_1: \mu_X - \mu_Y \neq 0$

accepts iff the $100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$ contains zero.

Soln

accept null iff

$$\left| \frac{\bar{x} - \bar{y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \right| < t_{n+m-2}^{\left(\frac{\alpha}{2}\right)}$$

$$\iff -t\left(\tfrac{\alpha}{2}\right) < \frac{\bar{x} - \bar{y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} < t\left(\tfrac{\alpha}{2}\right)$$

$$\iff -t\left(\tfrac{\alpha}{2}\right) S_p \sqrt{\tfrac{1}{n} + \tfrac{1}{m}} < \bar{x} - \bar{y} < t\left(\tfrac{\alpha}{2}\right) S_p \sqrt{\tfrac{1}{n} + \tfrac{1}{m}}$$

$$\iff \bar{x} - \bar{y} - t\left(\tfrac{\alpha}{2}\right) S_p \sqrt{\tfrac{1}{n} + \tfrac{1}{m}} < 0 < \bar{x} - \bar{y} + t\left(\tfrac{\alpha}{2}\right) S_p \sqrt{\tfrac{1}{n} + \tfrac{1}{m}}$$

$$\text{i.e. } 0 \in \bar{x} - \bar{y} \pm t\left(\tfrac{\alpha}{2}\right) S_p \sqrt{\tfrac{1}{n} + \tfrac{1}{m}}$$

Sec 11.2

We have assumed X, Y both have same variance

It we remove this assumption

then $\frac{S_X^2}{n} + \frac{S_Y^2}{m}$ is a natural

unbiased estimator of $\text{Var}(\bar{X} - \bar{Y})$

Chk $E\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right) = \frac{E(S_X^2)}{n} + \frac{E(S_Y^2)}{m}$

$= \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m} = \text{Var}(\bar{X}) + \text{Var}(\bar{Y})$

$= \text{Var}(\bar{X} - \bar{Y})$

X, Y indep

$\Rightarrow \frac{S_X^2}{n} + \frac{S_Y^2}{m}$ is unbiased estimator of $\text{Var}(\bar{X} - \bar{Y})$.

but what is

$$\frac{\bar{X} - \bar{Y} - 0}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \; ?$$

Answ it is approximately a t distribution w/ df given by formula on p428.

P428 Rice

We have used the assumption that the two populations have the same variance. If the two variances are not assumed to be equal, a natural estimate of $\mathrm{Var}(\overline{X} - \overline{Y})$ is

$$\frac{s_X^2}{n} + \frac{s_Y^2}{m}$$

If this estimate is used in the denominator of the t statistic, the distribution of that statistic is no longer the t distribution. But it has been shown that its distribution can be closely approximated by the t distribution with degrees of freedom calculated in the following way and then rounded to the nearest integer:

$$\mathrm{df} = \frac{[(s_X^2/n) + (s_Y^2/m)]^2}{\dfrac{(s_X^2/n)^2}{n-1} + \dfrac{(s_Y^2/m)^2}{m-1}}$$

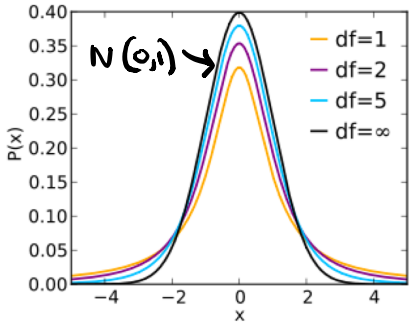
↖ no need to memorize, need to calculate this in R for HW.

Ex 50 rats were randomly divided into 2 grps of 25 each. The rats in one grp were given steroids. They were timed running a maze. The rats in the non steroid grp had an avg time of 10 seconds, with SD of 2 seconds. The rats in the steroid grp had an avg time of 9 seconds and SD=5.

a) What is the SE of difference of the 2 averages?

b) Calculate a test statistic for testing the null hypothesis that steroids make no difference.

c) What is the p-value of the test?
(using formula for df of t-test gives 31 df. This is large so just do a z-test).

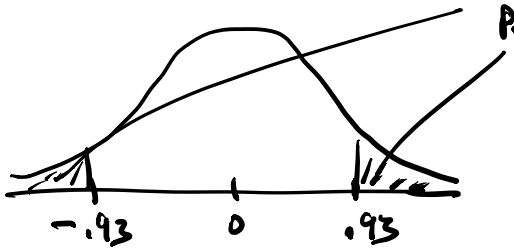


a) $s_{\bar{x}_T - \bar{x}_C} = \sqrt{\frac{s_T^2}{n_T} + \frac{s_C^2}{n_C}} = 1.08$

b) $H_0: \mu_T = \mu_C$
$H_1: \mu_T \neq \mu_C$

$T = \frac{\bar{x}_T - \bar{x}_C - 0}{1.08} = -.93$

c)



P-val $= P_0(|T| > |-.93|$
$= P_0(T < -.93$ or $T > .93)$
$= 2\Phi(-.93) = .35$

$\Phi(-.93) = P_{norm}(-.93) = .1768$

$.35 > \alpha = .05$

so don't reject the null.

Sec 11.2.2 Power

Calculation of power are an important part of planning an experiment in order to determine how large of a sample size you should use.

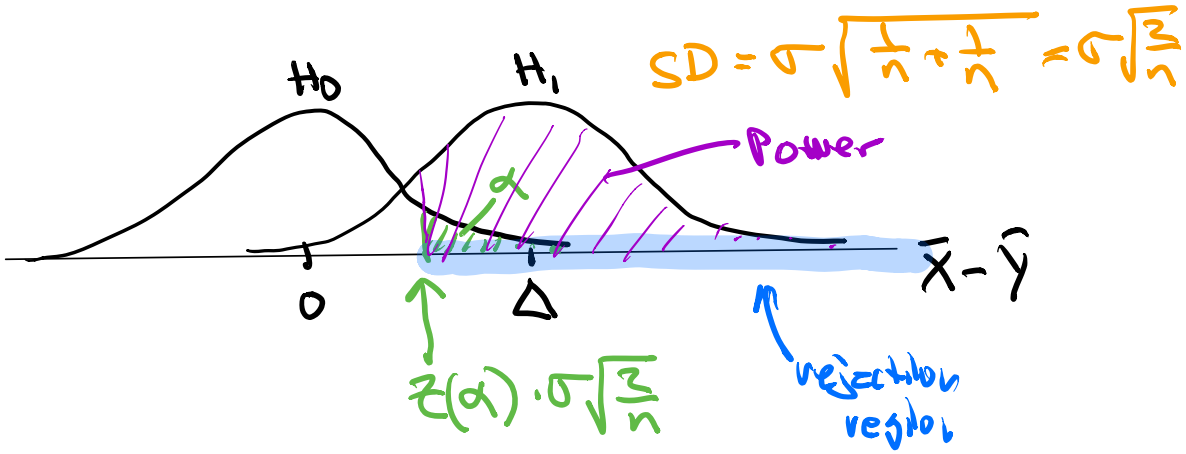
Let assume σ known
σ same in both populations
$n = m$

Picture

$H_0: \mu_x - \mu_y = 0$

$H_1: \mu_x - \mu_y = \Delta > 0$ — one sided alternative.

α level of significance.



$$\text{Power} = P_1\left(\phi(x)=1\right)$$

$$= P_1\left(\bar{x}-\bar{y} > z(\alpha)\sigma\sqrt{\tfrac{2}{n}}\right)$$

$$= P_1\left(\frac{\bar{x}-\bar{y}-\Delta}{\sigma\sqrt{\frac{2}{n}}} > \frac{z(\alpha)\sigma\sqrt{\frac{2}{n}}-\Delta}{\sigma\sqrt{\frac{2}{n}}}\right)$$

$$= \boxed{1-\Phi\left(z(\alpha)-\frac{\Delta}{\sigma\sqrt{\frac{2}{n}}}\right)}$$

As increase Δ, α, σ, n what happens to power?

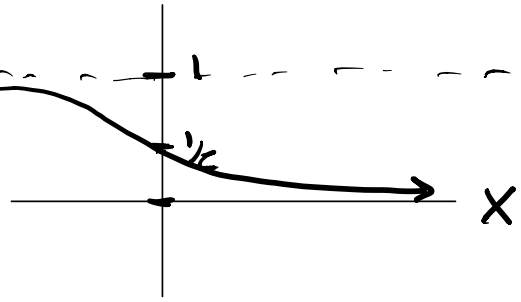
$\Delta \uparrow$ Power $\uparrow$

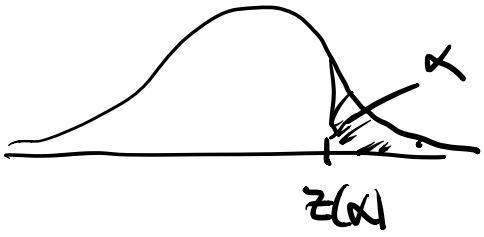
$\alpha \uparrow$ Power $\uparrow$

$\sigma \uparrow$ Power $\downarrow$

$n \uparrow$ Power $\uparrow$

Power $= 1 - \Phi(x)$





ex Suppose n measurements are taken from a treatment group and independently n are taken from a control group. An SD of a single measurement is $\sigma = 10$ for both groups.

How large should n be so that the test

$H_0 : \mu_x - \mu_y = 0$

$H_1 : \mu_x - \mu_y = \Delta > 0$

has power .7 if $\Delta = \mu_x - \mu_y = 1$ and $\alpha = .05$?

$$\text{Power} = 1 - \Phi\left(z(\alpha) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right)$$

$$\Rightarrow \Phi\left(z(\alpha) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right) = 1 - \text{Power} = .3$$

(Power $= .7$)

$$\text{qnorm}(.3) = -.524$$

$$\Rightarrow z(\alpha) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}} = -.524$$

$\alpha = .05 \Rightarrow z(\alpha) = 1.645$

$\Delta = 1$

$\sigma = 10$

$$\Rightarrow \sqrt{\frac{n}{2}} = (.524 + 1.645)\frac{10}{1}$$

$$\Rightarrow \boxed{n = 941}$$

Stat 135 lec 22

Last time

2-sample t-test:

Compared means of independent normal populations with different variances

$$s_{\bar{X}-\bar{Y}} = \frac{s_X^2}{n} + \frac{s_Y^2}{n}$$ standard error.

Today

sec 11.3 Paired 2 sample t-test

Paired and unpaired tests in R

pre 9.5 The multinomial distribution and Chi squared test

Sec 11.3.1 Paired t-test.

We wish to estimate $\mu_x - \mu_y$ but in paired studies our samples are no longer independent.

let $x_1, \dots, x_n$ iid $N(\mu_x, \sigma_x^2)$

$y_1, \dots, y_n$ iid $N(\mu_y, \sigma_y^2)$

$\sigma_{xy} = \text{Cov}(x_i, y_i) \neq 0.$

let $D_i = x_i - y_i$

$E(D_i) = \mu_x - \mu_y$ ← call this μ_D

$\text{Var}(D_i) = \text{Var}(x_i - y_i) = \sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}$

$$\bar{D} \approx N\left(\mu_x - \mu_y, \frac{1}{n}\left(\sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}\right)\right)$$

Notice If x, y are pos correlated

$\sigma_{xy} > 0$ and $\text{Var}(\bar{D}) < \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{n}$ ← $\text{Var}(\bar{x} - \bar{y})$ when x, y indep.

$\Rightarrow$ efficiency $= \frac{s_{\text{paired}}}{s_{\text{unpaired}}} < 1$

$\Rightarrow$ longer t.s. $\Rightarrow$ so more likely to reject the null for paired data.

$$t = \frac{\bar{D} - \mu_D}{s_{\bar{D}}} \sim t_{n-1}$$

$H_0: \mu_D = 0$
$H_1: \mu_D \neq 0$ has rejection region $|\bar{D}| > t_{n-1}\left(\frac{\alpha}{2}\right) s_{\bar{D}}$

lec 22 in-class exercise on b-courses

Stat135 lecture 22

Start Over

Examine your data

```
X <- c(24.6,17,16,10.4,8.2,7.9,8.2,7.9,5.8,5.4,5.1,4.7)
Y <- c(10.1,5.7,5.6,3.4,6.5,0.7,6.5,0.7,6.1,4.7,2.0,2.9)
n <- length(X)
df <-data.frame(X,Y)
df
```

X <dbl>	Y <dbl>
24.6	10.1
17.0	5.7
16.0	5.6
10.4	3.4
8.2	6.5
7.9	0.7
8.2	6.5
7.9	0.7
5.8	6.1
5.4	4.7

1-10 of 12 rows Previous 1 2 Next

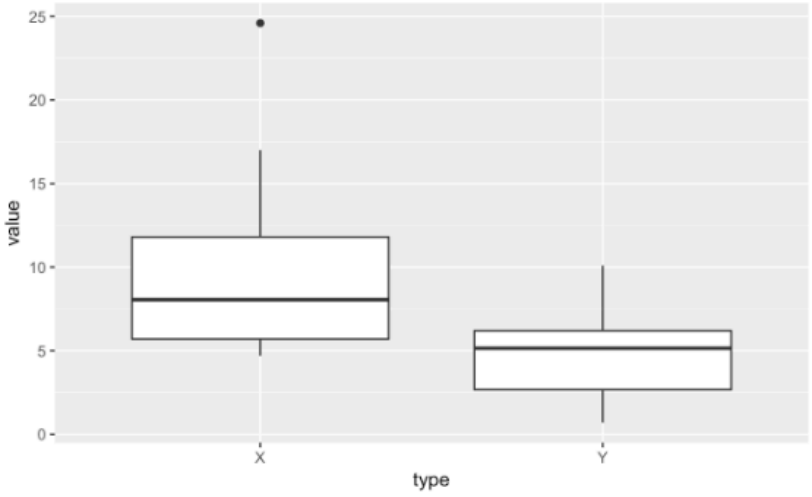
reorganize data so that it is "tidy"

```
library(tidyr)
#we reformat the data frame to tidy format for easier analysis
df_tidy<- df %>% gather(key=type, value=value, `X`, `Y`)
head(df_tidy)
```

	type <chr>	value <dbl>
1	X	24.6
2	X	17.0
3	X	16.0
4	X	10.4
5	X	8.2
6	X	7.9

6 rows

```
library(ggplot2)
df_tidy %>% ggplot(aes(x=type,y=value)) + geom_boxplot()
```



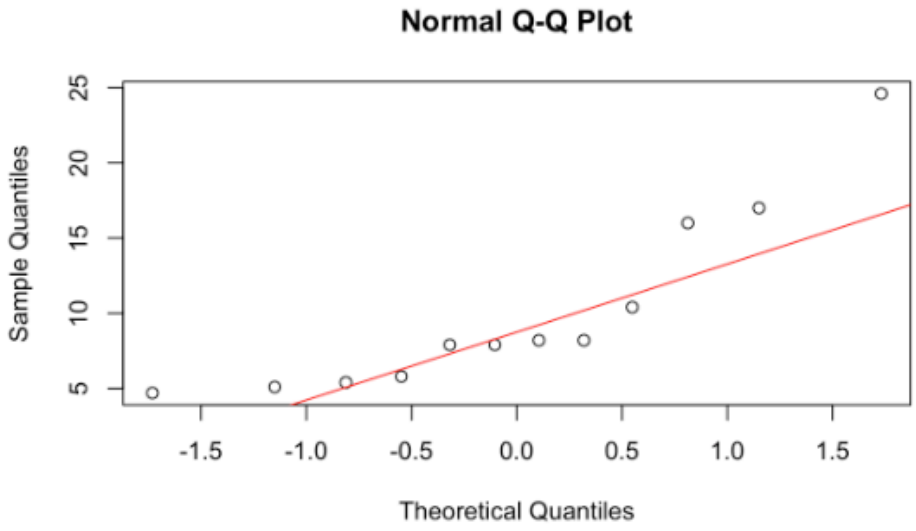
```
df_tidy
```

type <chr>	value <dbl>
X	24.6
X	17.0
X	16.0
X	10.4
X	8.2
X	7.9
X	8.2
X	7.9
X	5.8
X	5.4

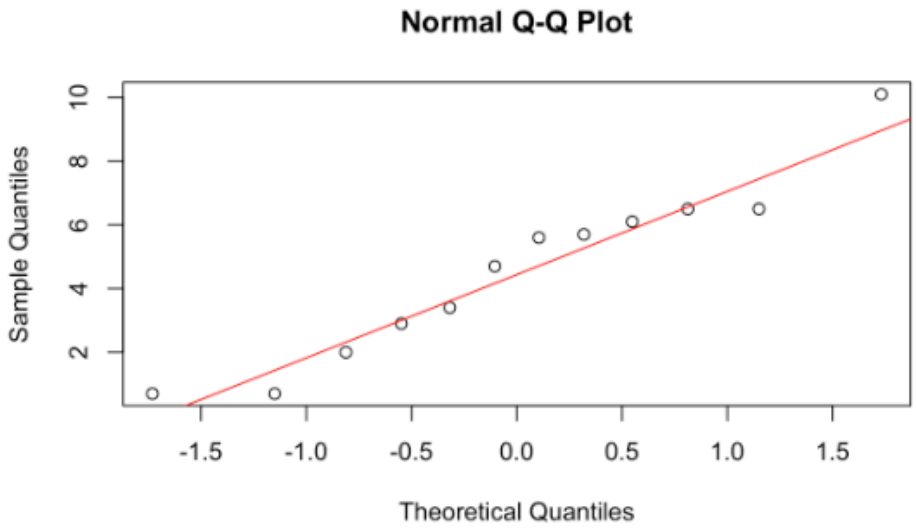
1-10 of 24 rows Previous 1 2 3 Next

Lets examine that the data is approximately normal and that the variances are approximately equal.

```
qqnorm(X); qqline(X,col=2)
```




```
qqnorm(Y); qqline(Y,col=2)
```



```
S.X <-  sd(X)  #sd() in R is sample SD
S.Y <-  sd(Y)
S.X
```

```
## [1] 5.80517
```

```
S.Y
```

```
## [1] 2.636957
```

Neither appear to be very normal and the variances don't appear to be the same. At this point I would do a nonparametric test called the Mann-Whitney test (after midterm). But lets do a parametric test and assume X and Y are normal but not assume the variances are the same.

Test that two independent samples come from populations with the same mean.

Since the variances aren't the same lets use the formula for degrees of freedom in the t-test on page 428 rounded to the nearest integer. Here n=m.

$$deg.freedom = \frac{((S_X^2/n) + (S_Y^2/n))^2}{(S_X^2/n)^2/(n-1) + (S_Y^2/n)^2/(n-1)}$$

```
deg.freedom <- round((n-1)*(S.X^2/n+S.Y^2/n)^2/((S.X^2/n)^2+(S.Y^2/n)^2))
deg.freedom
```

```
## [1] 15
```

$H_0: \mu_X - \mu_Y = 0$

$H_1: \mu_X - \mu_Y \neq 0$

$\alpha = .01$

```
t <- (mean(X)-mean(Y))/sqrt(S.X^2/n+S.Y^2/n)
t
```

```
## [1] 3.001744
```

```
p_val <- 2*(1-pt(abs(t),df=deg.freedom))
p_val
```

```
## [1] 0.008940922
```

p.val < .01 so reject null.

Test that two paired samples come from populations with the same mean.

We assume that before and after are normal vectors.

```
before <- c(24.6,17,16,10.4,8.2,7.9,8.2,7.9,5.8,5.4,5.1,4.7)
after <- c(10.1,5.7,5.6,3.4,6.5,0.7,6.5,0.7,6.1,4.7,2.0,2.9)
D <- before-after
S.D <- [illegible]sd(D)

D<- before-after
t <- (mean(D))/(S.D/sqrt(n))
t
```

bigger t value since SE for paired is smaller than for unpaired

```
## [1] 4.179141
```

```
p_val <- 2*(1-pt(abs(t),df=n-1))
p_val
```

smaller than p-val for unpaired test since t value bigger, This is good if alternative true, bad otherwise.

```
## [1] 0.001538782
```

p.val < .01 so reject the null. Notice that SE for the sampling distribution for the paired test is less then the sampling distribution for the unpaired test. In otherwords paired sampling is more efficient.

```
(SE_pair <- S.D/sqrt(n) ) #SE of paired sampling distribution
```

```
## [1] 1.322042
```

```
(SE_unpair <- sqrt((S.X^2/n)^2+(S.Y^2/n)^2)) #SE of unpaired sampling distribution
```

```
## [1] 2.867492
```

```
SE_pair/SE_unpair  #efficiencey
```

```
## [1] 0.4610447
```

< 1.

Sec 9.5 The multinomial distribution

Generalizes the binomial distribution

The multinomial RV is a sum of n indep m outcome (cells) RV.

($m=2$ for binomial.)

$$\left.\begin{array}{l} p_1 = \text{prob of outcome } 1 \\ p_2 = \text{" " " } 2 \\ \vdots \\ p_m = \text{prob of outcome } m \end{array}\right\} \quad p_1 + \cdots + p_m = 1$$

What is chance of getting x_1 outcome 1, x_2 outcome 2, ... x_m outcome m? $(n = x_1 + \cdots + x_m)$

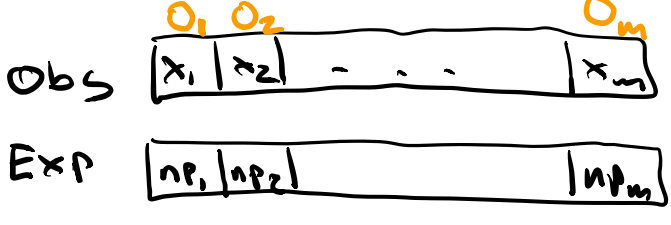
Answ $\binom{n}{x_1, x_2, \ldots, x_m} p_1^{x_1} \cdots p_m^{x_m}$ multinomial formula,

$\binom{n}{x_1, x_2, \ldots, x_m} = \frac{n!}{x_1! \cdots x_m!}$

If we know $p_1, \ldots, p_m$ the multinomial distribution is completely specified.

(i.e. we can calculate the prob of any outcome)

In n trials



$\sum_{i=1}^{m} x_i = n$

$\sum_{i=1}^{m} p_i = 1$

$X_i \sim \text{Bin}(n, p_i) \approx N(np_i, np_iq_i)$

Chi square Statistic $\sum_{i=1}^{m \leftarrow \#\text{cells}} \frac{(O_i - E_i)^2}{E_i} \xrightarrow{n\to\infty} \chi^2_{m-1}$

When we have 2 outcomes (binomial)
we will show this is χ^2_1

$$\sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} = \frac{(x_1 - np_1)^2}{np_1} + \frac{(x_2 - np_2)^2}{np_2}$$

$E(x_1) = np_1$, $\mathrm{Var}(x_1) = np_1q_1$

algebra $\rightarrow$ $\frac{(x_1 - np_1)^2}{np_1q_1} = z^2 \sim \chi^2_1$

More generally (algebra; for large n)

$$\sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{m-1} \frac{(O_i - E_i)^2}{E_i q_i} = \sum_{i=1}^{m-1} z_i^2 \sim \chi^2_{m-1}$$

Ex Please review the example of modeling the arrival of α particles in a Poisson random scatter in lecture 3 (intro to chap 8).

We showed the table below which is the observed and expected number of arrivals in 1207 10 second intervals.

We model the number of arrivals in a 10 second interval as $X \sim \text{Pois}(\lambda \cdot 10)$. We estimated $\hat{\lambda} = .84$ so $\lambda \cdot 10 = 8.4$ arrivals/10 sec.

Here $P_1 = e^{-8.4} + e^{-8.4}(8.4) + \frac{e^{-8.4}(8.4)^2}{2!}$

$P_2 = \frac{e^{-8.4}(8.4)^3}{3!} = .022$

etc.

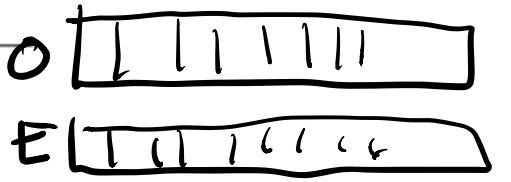
Expected count for $n=3$ arrivals is $1207(.022) = 27$

Think of a box with 16 outcomes with probabilities $P_1, P_2, \ldots, P_{16}$

$[P_1, P_2, P_3 \ldots \ldots P_{16}]$

↓ 1207 draws

	n	Observed	Expected
P_1	0–2	18	12.2
P_2	3	28	27.0
P_3	4	56	56.5
.	5	105	94.9
.	6	126	132.7
.	7	146	159.1
.	8	164	166.9
.	9	161	155.6
.	10	123	130.6
.	11	101	99.7
.	12	74	69.7
.	13	53	45.0
.	14	23	27.0
.	15	15	15.1
.	16	9	7.9
P_{16}	17+	5	7.1
		1207	1207



↖ these bars are the table

Null: Our observed counts come from the multinomial distribution $MN(1207, P_1(\lambda), P_2(\lambda) \ldots, P_{16}(\lambda))$ where $P_i(\lambda)$ is Poisson (λ).

Alt: Our observed counts come from some other multinomial distribution

we calculate χ^2 stat

$$\sum_{\text{all rows}} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{16-1-1}$$

- $16-1$: dim of sample space $(p_1, \ldots, p_{16})$ where $\sum_{i=1}^{16} p_i = 1$
- -1: dim of null space (λ is 1 unknown parameter in Pois(λ))

Stat 135 lec 23

Last time

Sec 9.5 goodness of fit χ^2 test.

In chap 11 we did 1,2 sample t test for the mean of a normal RV. Our box had a continuum of tickets w/ mean μ, var σ^2.

Now we have a categorical RV having m outcomes having probabilities $p_1, \ldots, p_m$. The chance a sample of size n for our box has a certain composition is given by the multinomial formula $P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) = \binom{n}{x_1, \ldots, x_m} p_1^{x_1} \cdots p_m^{x_m}$

The goal is to test whether a model for the population distribution $(p_1, \ldots, p_m)$ fits our data.

We draw n times with replacement from our box and get observed counts $\boxed{x_1 | x_2 | \cdots | x_m}$ $x_1 + \cdots + x_m = n$

If the prob of tickets in box is $p_1(\theta), \ldots, p_m(\theta)$ we expect to get counts $\boxed{np_1(\theta) | np_2(\theta) | \cdots | np_m(\theta)}$

We do a goodness of fit test.

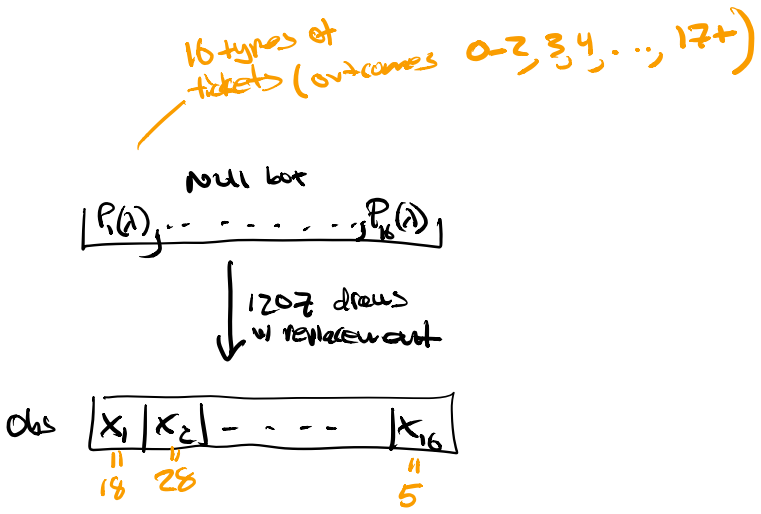
Pearson Chi Square T.S. $\sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{m-1-k}$ where k is the dimension of θ.

A goodness of fit test explores how good your probability model fits your data,

If the p-value of our Pearson Chi Square T.S. is smaller than α we reject the null hypothesis.

ex

n	Observed	Expected
0–2	18	12.2
3	28	27.0
4	56	56.5
5	105	94.9
6	126	132.7
7	146	159.1
8	164	166.9
9	161	155.6
10	123	130.6
11	101	99.7
12	74	69.7
13	53	45.0
14	23	27.0
15	15	15.1
16	9	7.9
17+	5	7.1
	1207	1207



Null: Our observed counts come from the multinomial distribution $MN(1207, P_1(\lambda), P_2(\lambda) \ldots, P_{16}(\lambda))$ where $P_i(\lambda)$ is Poisson(λ).

Alt: Our observed counts come from some other multinomial distribution

we calculate χ^2 stat

$$\sum_{\text{all cells}} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{16-1-1}$$

$16-1$: $M-1$; -1: K

Compare P value to $\alpha = .05$.

Today

Sec 9.5 GLRT for the multinomial distribution,

① Example

② theory: we will see $-2\log\Lambda \approx \sum_{i=1}^{m \leftarrow \#\text{cells}} \frac{(O_i - E_i)^2}{E_i}$

we know $-2\log\Lambda \sim \chi^2_{\dim\Omega - \dim\omega_0}$ (Ω: sample space, ω_0: null space)

$$\Rightarrow \sum_{i=1}^{m \leftarrow \#\text{cells}} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{\dim\Omega - \dim\omega_0}$$

(1) Hardy Weinberg (HW) equilibrium model

If gene frequencies are in equilibrium the genotypes AA, Aa, aa occur in the population w/ prob $(1-\theta)^2$, $2\theta(1-\theta)$, θ^2 according to the HW model.

Null box $\boxed{P_1(\theta)=(1-\theta)^2 \mid P_2(\theta)=2\theta(1-\theta) \mid P_3(\theta)=\theta^2}$

We observe blood type

	M (AA)	MN (Aa)	N (aa)	
Obs	342	500	187	$n=1029$
	x_1	x_2	x_3	

Step 1 Find $\hat{\theta}_{ML}$

$$lik(\theta) = \frac{n!}{x_1!x_2!x_3!}(1-\theta)^{2x_1}(2\theta(1-\theta))^{x_2}\theta^{2x_3}$$

$$l(\theta) = \log n! - \sum_{i=1}^{3}\log x_i! + x_1\log(1-\theta)^2 + x_2\log 2\theta(1-\theta) + x_3\log\theta^2$$

$$l'(\theta)=0 \Rightarrow \hat{\theta}_{ML} = \frac{2x_3+x_2}{2n} = \frac{2(187)+500}{2(1029)} = \boxed{.4277}$$

So

$$P_1(\hat{\theta}) = (1-.4277)^2 = .3200$$

$$P_2(\hat{\theta}) = 2(.4277)(1-.4277) = .4887$$

$$P_3(\hat{\theta}) = (.4277)^2 = .1804$$

Step 2

Exp

$nP_1(\hat{\theta})$	502.83	185.60

$= 340.57$

Test

H_0: have $MN(1029, .32, .49, .18)$

H_1: have some other MN w/ $n = 1029$.

O

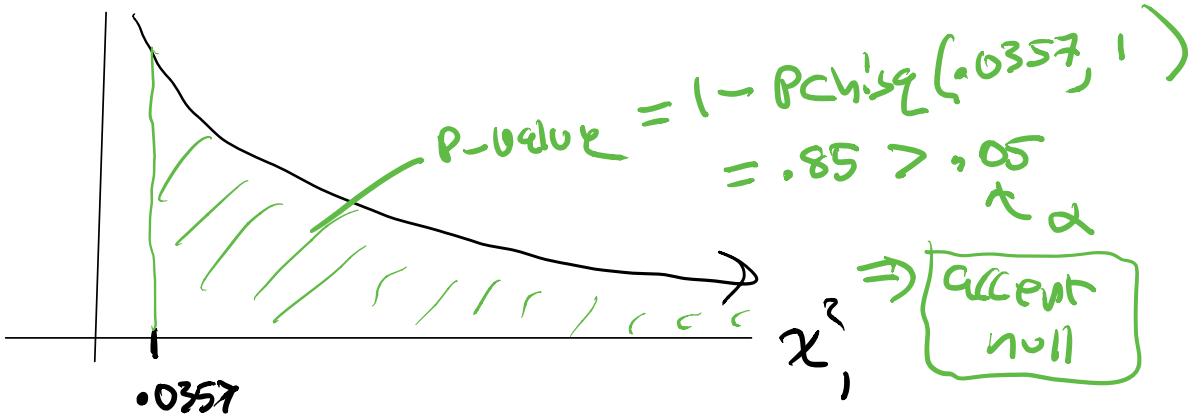
342	500	187

E

340.4	502.8	183.6

$$\chi^2_{\underbrace{3-1-1}_{1}} = \frac{(342 - 340.6)^2}{340.6} + \cdots$$

$$= .0357$$



② Sec 9.5 Theory.

H_0: have a MN distribution of cell prob

$p(\theta) = (p_1(\theta), \ldots, p_m(\theta))$ where $\theta = (\theta_1, \ldots, \theta_k) \in \omega_0 \subseteq \mathbb{R}^k$ (open interval)

H_1: have a MN dist w/ different cell prob than our null.

Ω = set of m nonneg numbers that sum to 1 $(p_1, \ldots, p_m)$ w/ $p_1 + p_2 + \cdots + p_m = 1$

Sample Space.

$$\Lambda = \frac{\max_{\theta \in \omega_0} \left(lik(p_1(\theta), \ldots, p_m(\theta)) \right) = lik(p_1(\hat{\theta}), \ldots, p_m(\hat{\theta}))}{\max_{(p_1, \ldots, p_m) \in \Omega} \left(lik(p_1, \ldots, p_m) \right) = lik(\hat{p}_1, \ldots, \hat{p}_m)}$$

Fact $\hat{p}_i = \frac{x_i}{n}$ ← see Sec 8.5.1 p273 Use Lagrange Multipliers.

$$\boxed{x_i = n\hat{p}_i}$$

By consistency of MLE

$$P_i(\hat{\theta}) \longrightarrow P_i(\theta)$$

and $\hat{P}_i \longrightarrow P_i(\theta)$ if null is true

$$\Lambda = \frac{\frac{n!}{x_1! \cdots x_m!} P_1(\hat{\theta})^{x_1} \cdots P_m(\hat{\theta})^{x_m}}{\frac{n!}{x_1! \cdots x_m!} \hat{P}_1^{x_1} \cdots \hat{P}_m^{x_m}}$$

$$= \prod_{i=1}^{m} \left(\frac{P_i(\hat{\theta})}{\hat{P}_i} \right)^{x_i}, \quad x_i \leftarrow n\hat{P}_i$$

$$-2 \log \Lambda = -2 \sum_{i=1}^{m} (n\hat{P}_i) \log \left(\frac{nP_i(\hat{\theta})}{n\hat{P}_i} \right), \quad n\hat{P}_i = O_i, \quad nP_i(\hat{\theta}) = E_i$$

$$= \boxed{2 \sum_{i=1}^{m} O_i \log \left(\frac{O_i}{E_i} \right)}$$

Fact (see P342 (Taylor series argument))

$$2 \sum_{i=1}^{m} O_i \log \left(\frac{O_i}{E_i} \right) \approx \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

↖ Pearson χ^2 statistic

$\dim \Omega = m-1$

$\omega_0 = \{(\theta_1, \ldots, \theta_k)$ s.t. θ_i is in an open interval in $\mathbb{R}\}$

$\dim \omega_0 = k.$

$$\Rightarrow -2\log\lambda \longrightarrow \chi^2_{m-1-k}$$

$$\| \quad \| \quad \| \quad \| \quad \|$$

$$\sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

Stat 135 lec 24

Last time

Sec 9.5 Application of GLRT to goodness of fit χ^2 test.

The goodness of fit (G.O.F) χ^2 test tests whether a categorical random variable follows a certain distribution (null hypothesis), versus that it follows some other distribution.

We saw that $-2\log\Lambda$, where Λ is from the GLRT (see theory in lec 23), is approximately the Pearson χ^2 test statistic, $\sum_{i=1}^{\#\text{cells}} \frac{(O_i - E_i)^2}{E_i}$.

Since $-2\log\Lambda \sim \chi^2_{(\dim \text{sample space} - \dim \text{null space})}$, it follows that $\sum_{i=1}^{\#\text{cells}} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(\dim \text{sample space} - \dim \text{null space})}$.

Today

(1) Assumptions of goodness of fit test,

(2) Examples.

(1) Assumptions of goodness of fit test

(0) have one categorical variable

(1) have indep observations
(draw w/ replacement)

(2) The options are mutually exclusive.
(i.e. every ticket goes in exactly one cell)

(3) we require large n and no more than 20% of expected counts < 5

small E_i and large n then

$E_i = np_i \Rightarrow p_i$ small

$X_i \sim Bin(n, p_i)$

want

$X_i \sim N(np_i, np_i z_i)$

but for small p_i, $X_i \sim Poisson(\mu = np_i)$

not normal,

(2) Examples of GOF test

ex

As part of a study on the selection of grand juries in Alameda county, the educational level of grand jurors was compared with the county distribution:[14]

Educational level	*County*	*Number of jurors*
Elementary	28.4%	1
Secondary	48.5%	10
Some college	11.9%	16
College degree	11.2%	35
Total	100.0%	62

Could a simple random sample of 62 people from the county show a distribution of educational level so different from the county-wide one?

Solution

As part of a study on the selection of grand juries in Alameda county, the educational level of grand jurors was compared with the county distribution:[14]

Educational level	*County*		*Number of jurors*
Elementary	$p_1 = 28.4\%$	$E_1 = np_1$	1
Secondary	$p_2 = 48.5\%$	$E_2 = np_2$	10
Some college	$p_3 = 11.9\%$	$E_3 = np_3$	16
College degree	$p_4 = 11.2\%$	$E_4 = np_4$	35
Total	100.0%		62 $= n$

Could a simple random sample of 62 people from the county show a distribution of educational level so different from the county-wide one?

H_0 : multinom(62, $p_1 = 28.4$, $p_2 = 48.5$, $p_3 = 11.9$, $p_4 = 11.2$)

H_1 : some other multinomial dist.

O	E
1	17.6
10	29.8
16	7.4
35	6.9

$$\chi^2_3 = \frac{(1-17.6)^2}{17.6} + \frac{(10-29.8)^2}{29.8} + \frac{(16-7.4)^2}{7.4} + \frac{(35-6.9)^2}{6.9}$$

$$= 153.24$$

$$\chi^2_3(.05) = 7.814$$

↖ qchisq(.95, df=3)

$\Rightarrow$ reject null

ex

(10 pts) Suppose the school claims that the number of incoming students for the past 4 years has been steadily increasing by 10% per year. Assuming that there is no dropout, form a null and alternative hypothesis around this claim. Name the test to be carried out, calculate the degrees of freedom, and list out the observed values (O) and their corresponding expected values (E).
Note: You do NOT need to calculate the final result or perform the test.

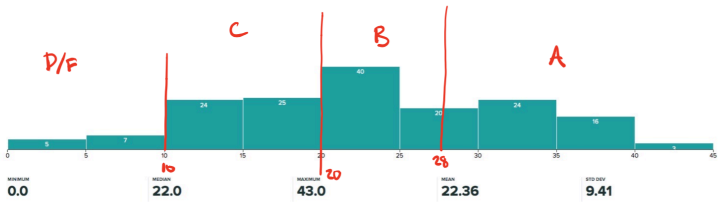
try and solve this.

) (10 pts) Suppose the school claims that the number of incoming students for the past 4 years has been steadily increasing by 10% per year. Assuming that there is no dropout, form a null and alternative hypothesis around this claim. Name the test to be carried out, calculate the degrees of freedom, and list out the observed values (O) and their corresponding expected values (E).
Note: You do NOT need to calculate the final result or perform the test.
Solution:
We perform a Goodness of Fit test.
H_0 : The number of students in each year is increasing by 10% per year.
H_1 : The number of students in each year is not as described.

Since we are assuming that there are no students dropping out, the null hypothesis states that the number of students in each year in our sample follows a multinomial distribution with probabilities equal to the proportion of students in each year. To calculate the exact null distribution, suppose the proportion of seniors in Berkeley is p, then under the null the probabilities of a randomly selected student being a Freshman, Sophomore, Junior and Senior are $(p * 1.1^3, p * 1.1^2, p * 1.1, p)$, respectively. Under the constraint that the sum of these 4 probabilities is 1, we can solve for $p : p \approx 0.215$. Plugging in p and the total sample size ($n = 904$), the expected numbers of students in each year are $904 * (p * 1.1^3, p * 1.1^2, p * 1.1, p) = (259.26, 235.69, 214.26, 194.79)$ and our observed values are $(241, 227, 216, 220)$. The degrees of freedom for this test is $4 - 1 = 3$

Next time, test of homogeneity and independance,
Sec 13.3, 13.4

Stat 135 lec 25



Nice job on a tough midterm!

Last time sec 9.5 goodness of fit.

Test that one categorical variable has certain multinomial distribution with cell prob $p_1(\theta), \dots, p_m(\theta)$ where $\theta = (\theta_1, \dots, \theta_k)$ and $p_1(\theta) + \cdots + p_m(\theta) = 1$.

Key result: $$\sum_{\#\text{cells}} \frac{(O_i - E_i)^2}{E_i} \approx \chi^2_{\dim \Omega - \dim \omega_0}$$

$\dim \Omega = m-1$, $\dim \omega_0 = k$

Today

sec 13.3, 13.4 Test of Homogeneity and Independence

1 cat variable
- GOF — does the cat var. have the claimed multivariate distribution
- test of homogeneity — do 2 or more subgps of a pop share the same multivariate dist.

2 cat variables — test of indep — are the 2 cat variables independent?

Sec 13.3 Test of homogeneity

ex 13.3.1

A study was done comparing frequencies of a particular allele in a sample of diabetics and non diabetics

Data was observed:

I \ J	Diab	non Diab	tot
Bb or bb	12	4	16
BB	39	49	88
	51	53	104

Are the relative frequency of the alleles significantly different in the 2 grps Diab, non Diab

Theory

J indep, I cell multinomials

H_0: all J multinomials are same

H_1: not all the same

Picture

I = 2

J = 3

null bd $\begin{bmatrix}\pi_1\\ \pi_2\end{bmatrix} = \begin{bmatrix}\pi_1\\ \pi_2\end{bmatrix} = \begin{bmatrix}\pi_1\\ \pi_2\end{bmatrix}$

$n_{\cdot 1}$ $n_{\cdot 2}$ $n_{\cdot 3}$ → J

$\hat{\pi}_1 = \frac{n_{1\cdot}}{n}$

$\hat{\pi}_2 = \frac{n_{2\cdot}}{n}$

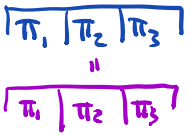
obs

I				
	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n

Exp

What if we divide population according to allele type instead of according to whether or not you have diabetes? You get the same Obs and Exp tables!!

Notice if you have I indep, J cell multinomials



Null Hx

$\hat{\pi}_1 = \frac{n_{\cdot 1}}{n}$

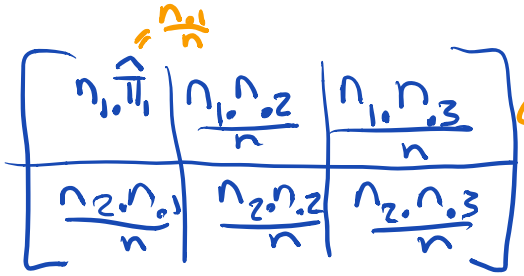
$\hat{\pi}_2 = \frac{n_{\cdot 2}}{n}$

$\hat{\pi}_3 = \frac{n_{\cdot 3}}{n}$

Obs

← same as before

Exp



← Same as before

$$\chi^2_{df} = \dim \Omega - \dim \omega_0$$

ω_0 = have 1 multinomial (null hyp)

w/ $I-1$ free params

$\dim \omega_0 = I-1$

Ω = have J multinomials in group

w/ $I-1$ free params per multinomial

$\Rightarrow \dim \Omega = J(I-1)$

$\Rightarrow df = \dim \Omega - \dim \omega_0 = J(I-1) - (I-1)$

$= \boxed{(I-1)(J-1)}$

T.S $\chi^2_{(I-1)(J-1)} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

$= \sum_{i,j} \left(\frac{\left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n} \right)^2}{\frac{n_{i\cdot} n_{\cdot j}}{n}} \right)$

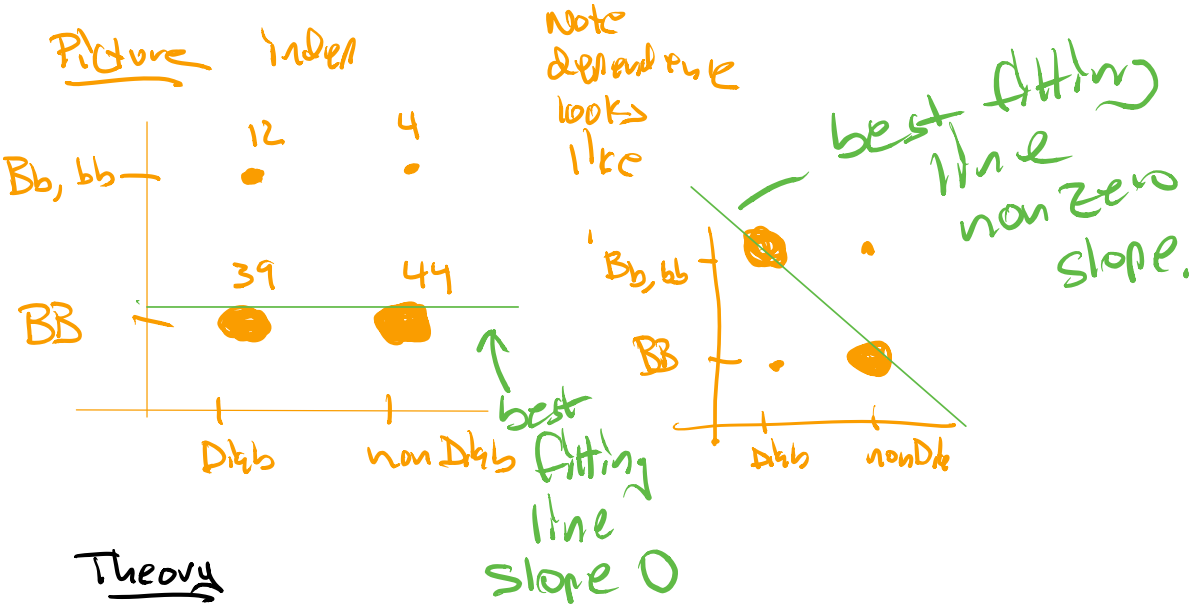
Sec 13.4 Test of Indep

- test whether 2 cat variables are indep.

⇐ modify Problem 13.8.1

Same data but diff question.

Is there a relationship between whether you have diabetes and your allele type?



Theory

1 J cell multinomial

1 I cell multinomial

notation

$$P(J=j) = \Pi_{J=j} = \Pi_{j}$$

$$P(I=i) = \Pi_{I=i} = \Pi_{i}$$

$$P(I=i, J=j) = \Pi_{I=i, J=j} = \Pi_{ij}$$

Indep means $\Pi_{ij} = \Pi_{i} \Pi_{j}$

H_0: two RV are indep $\left(\text{i.e } \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad \forall i,j\right)$

H_1: two RV are dep $\left(\text{i.e } \pi_{ij} \neq \pi_{i\cdot}\pi_{\cdot j} \text{ some } i,j\right)$

Picture

$I = 2$, $J = 3$

	π_1	π_2	π_3
π_1	$\pi_1\pi_1$	$\pi_1\pi_2$	$\pi_1\pi_3$
π_2	$\pi_2\pi_1$	$\pi_2\pi_2$	$\pi_2\pi_3$

null box

$\downarrow n$

Obs

$$\begin{bmatrix} n_{11} & n_{12} & n_{13} \\ n_{21} & n_{22} & n_{23} \end{bmatrix} \begin{matrix} n_{1\cdot} \\ n_{2\cdot} \end{matrix}$$

$$n_{\cdot 1} \quad n_{\cdot 2} \quad n_{\cdot 3}$$

$$\hat{\pi}_{I=1} = \frac{n_{1\cdot}}{n}$$

$$\hat{\pi}_{J=1} = \frac{n_{\cdot 1}}{n}$$

Same T.S !!

We will show next time that $df = (I-1)(J-1)$ so test of independance and test of homogeneity are equivalent tests for 2 different experimental designs.

Stat 135 Sec 76

Last time Sec 9.5 Sec 13.3, 13.4

Goodness of Fit — single sample from population
Answers Q: Does population have a particular multinomial distribution?

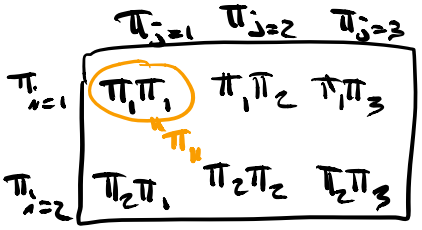
Test of Homogeneity — multiple samples from subgroups of a single population
Answers Q: Do subgroups have same multinomial dist?

Test of Independence — single sample from population
Answers Q: Are two categorical variables independent?

Test of independence H_0: two Cat variables indep
H_1: dependent.

Picture

$I=2$
$J=3$



null box

$\downarrow n$

obs $\begin{bmatrix} n_{11} & n_{12} & n_{13} \\ n_{21} & n_{22} & n_{23} \\ n_{\cdot 1} & n_{\cdot 2} & n_{\cdot 3} \end{bmatrix}$ $\begin{matrix} n_{1\cdot} \\ n_{2\cdot} \\ \end{matrix}$

Exp $\begin{bmatrix} \frac{n_{1\cdot} n_{\cdot 1}}{n} & \frac{n_{1\cdot} n_{\cdot 2}}{n} & \frac{n_{1\cdot} n_{\cdot 3}}{n} \\ \frac{n_{2\cdot} n_{\cdot 1}}{n} & \frac{n_{2\cdot} n_{\cdot 2}}{n} & \frac{n_{2\cdot} n_{\cdot 3}}{n} \end{bmatrix}$ $\leftarrow n\hat{\pi}_1\hat{\pi}_1$

Same tables as in test of Homogeneity.

Today

1. d.f. of Test of independence (Sec 13.4)
2. in class exercise
3. non parametric test version of unpaired t-test (Sec 11.2.3)
   "The Mann Whitney Test".

① d.f. of test of Independence (Sec 13.4)

ω_0 = dim of marginal prob of null box

$\pi_{i\cdot}$ $i=1,\dots,I$ $\quad$ $\pi_{\cdot j}$ $j=1,\dots,J$

w/ constraints $\sum_{i=1}^{I}\pi_{i\cdot}=1$, $\sum_{j=1}^{J}\pi_{\cdot j}=1$

$\Rightarrow$ dim $\omega_0 = I-1+J-1$

$$\Omega = \left\{ \pi_{ij} \;\; \begin{matrix} i=1,\dots,I \\ j=1,\dots,J \end{matrix} \;\; \text{w/} \sum_{ij}\pi_{ij}=1 \right\}$$

dim $\Omega = IJ-1$

$\Rightarrow$ df = dim Ω − dim ω_0

$= IJ-1-\left((I-1)+(J-1)\right) = \boxed{(I-1)(J-1)}$

Summary

Test of Homo and Indep have the same T.S, and df and differ by design of experiment.

In test of Indep, observational units are collected at random from the box and 2 categorical variables are observed for each unit.

In test of Homo, the data is collected by random sampling from each subgrp of the subpopulation seperately.

Assumptions of χ^2 tests (GOF, Homo, Indep)

① have one or two categorical variable

② have indep observations (i.e draw w/ replacement)

Note SRS or if relatively small sample size.

③ Outcomes are mutually exclusive (every ticket in exactly one cell)

④ We require n to be large and no more than 20% of expected counts < 5.

n large since $-2\log\Lambda \longrightarrow \chi^2_{\dim\Omega-\dim\omega_0}$

If E_{ij} is small and n is large

$E_{ij} = np_{ij} \Rightarrow p_{ij}$ small

$X_{ij} \sim Bin(n, p_{ij})$

want $X_{ij} \sim N(np_{ij}, np_{ij}q_{ij})$

but if p_{ij} is too small, $X_{ij} \sim Poisson(\mu = np_{ij})$

not normal.

(2) In-class exercise

For each question answer

1) What test to use
2) Null and alternative hypothesis
3) Are assumptions met
4) Degrees of freedom
5) Solve using chisq.test() in R

A)

9. This problem considers some more data on Jane Austen and her imitator (Morton 1978). The following table gives the relative frequency of the word *a* preceded by (PB) and not preceded by (NPB) the word *such*, the word *and* followed by (FB) or not followed by (NFB) *I*, and the word *the* preceded by and not preceded by *on*.

Words	*Sense and Sensibility*	*Emma*	*Sanditon* I	*Sanditon* II
a PB *such*	14	16	8	2
a NPB *such*	133	180	93	81
and FB *I*	12	14	12	1
and NFB *I*	241	285	139	153
the PB *on*	11	6	8	17
the NPB *on*	259	265	221	204

B) Was Austen consistent in these habits of style from one work to another? Did her imitator successfully copy this aspect of her style?

16. A market research team conducted a survey to investigate the relationship of personality to attitude toward small cars. A sample of 250 adults in a metropolitan area were asked to fill out a 16-item self-perception questionnaire, on the basis of which they were classified into three types: cautious conservative, middle-of-the-roader, and confident explorer. They were then asked to give their overall opinion of small cars: favorable, neutral, or unfavorable. Is there a relationship between personality type and attitude toward small cars? If so, what is the nature of the relationship?

	Personality Type		
Attitude	Cautious	Midroad	Explorer
Favorable	79	58	49
Neutral	10	8	9
Unfavorable	10	34	42

C)

To test whether a die is fair, someone rolls it 600 times. On each roll, he just records whether the result was even or odd, and large (4, 5, 6) or small (1, 2, 3). The observed frequencies turn out as follows:

	Large	*Small*
Even	183	113
Odd	88	216

Question: Is the die fair?

A

Test of Homogeneity since independent sample from each subgroup of population. There are a couple parts to the question the first one is whether Austin was consistent in her style. For this we will analyze the first 3 columns of the table to see whether they have the same multinomial distribution.

Null: each of the 3 multinomial distributions are the same

Alt: Not all the multinomial distributions are the same

Assumptions are met:
1)Each ticket is multinomial with one cell checked off
2)Not clear if this is a random sample of frequency of words but any case ok.
3) check expectation table

First we set up the data:

```
data <- c(14 , 16 , 8 , 133 , 180 , 93, 12 , 14 , 12 , 241 , 285 , 139 , 11 , 6 , 8, 259 , 265 , 221)
data.matrix <- matrix(data, nrow = 6, ncol = 3, byrow = TRUE)
data.matrix
```

```
##      [,1] [,2] [,3]
## [1,]   14   16    8
## [2,]  133  180   93
## [3,]   12   14   12
## [4,]  241  285  139
## [5,]   11    6    8
## [6,]  259  265  221
```

Now perform the test (first displaying the expected table for fun):

```
chisq.test(data.matrix, correct = FALSE)$expected
```

```
##            [,1]       [,2]       [,3]
## [1,]  13.281168  15.184142   9.534690
## [2,] 141.898800 162.230569 101.870631
## [3,]  13.281168  15.184142   9.534690
## [4,] 232.420449 265.722483 166.857068
## [5,]   8.737611   9.989567   6.272822
## [6,] 260.380803 297.689098 186.930099
```

```
chisq.test(data.matrix, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data.matrix
## X-squared = 23.287, df = 10, p-value = 0.009735
```

Thus, we reject the null; Austen was not consistent among her works.

To see if Austen's imitator was successful in imitating her style, we can first compile all of Austen's data into one vector (other wise it isn't quite clear which work to compare the imitator to since Austin wasn't consistent in her habits of style). We compare with Sanditon 2 as an example:

```
austen <- c()        # initialize an empty vector
for(i in 1:6){
  austen[i] = sum(data.matrix[i, ])
}
imitator <- c(2, 81, 1, 153, 17, 204)
matrix <- matrix(c(austen, imitator), nrow = 6, ncol = 2, byrow = FALSE)
chisq.test(matrix, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  matrix
## X-squared = 29.726, df = 5, p-value = 1.67e-05
```

Thus, once again, we reject the null; the imitator was unsuccessful in imitating Austen's style.

B)

Test of independence since a single SRS of 250.

Null: two categorical variables personality type and attitude are independent

Alt: dependent

Assumptions:
1) every ticket is the cross classification of two mutinomials with one cell checked off.
2) Every ticket is independent
3) check expectation

```
b <- matrix(c(79,58,49,10,8,9,10,34,42),nrow=3,ncol=3,byrow = TRUE)
b
```

```
##      [,1] [,2] [,3]
## [1,]   79   58   49
## [2,]   10    8    9
## [3,]   10   34   42
```

```
chisq.test(b)$expect
```

```
##           [,1]     [,2]     [,3]
## [1,] 61.585284 62.20736 62.20736
## [2,]  8.939799  9.03010  9.03010
## [3,] 28.474916 28.76254 28.76254
```

df=2*2=4

```
chisq.test(b)
```

```
##
##  Pearson's Chi-squared test
##
## data:  b
## X-squared = 27.289, df = 4, p-value = 1.737e-05
```

we reject the null that variables are independent.

C)

Tricky, looks like test of independence but it is a goodness of fit test since test of independance doesn't answer the quesiton about fairness. We have

Evan and Large= 4,6 —183 Evan and Small= 2 —-113 Odd and Large=5 —-88 Odd and Small= 1,3 —-216

Expect probabilities 2/6,1/6,1/6,2/6

Null: fair die

Alt: unfair die

Assumptions:
1) Every ticket is a multinomial with 4 cells with one cell checked off 2) tickets are independent since rolling die is indep 3) Like drawing with replacement from a box 4)expected table:

```
c <-  c(183,113,88,216)
c
```

```
## [1] 183 113  88 216
```

```
chisq.test(c,p=c(2/6,1/6,1/6,2/6))$expect
```

```
## [1] 200 100 100 200
```

df=3

```
chisq.test(c,p=c(2/6,1/6,1/6,2/6))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c
## X-squared = 5.855, df = 3, p-value = 0.1189
```

~~Odie~~ die is ~~not~~ fair, we ~~reject~~ accept the null.

③ Mann-Whitney test (Sec 11.2.3)

In section 11.2 (unpaired 2 sample t-test) we are comparing 2 independent samples of <u>normal data</u>.

If normality doesn't hold for small sample size use nonparametric test, called

- Mann-Whitney test
- Rank sum test
- Wilkoxin test in R

} all the same test.

lec 27

Q3 Friday 60F sec 9.5, Homogeneity, Independence Sec 13.3 13.4

Today
sec 11.2 Mann-Whitney test —— non parametric test version of unpaired t-test,
a.k.a. Wilcoxon - Rank Sum test

Motivation
We want assume normality of our data.

Notation
X small grp $\sim F$ (usually control) $x_1, \dots, x_n$
Y larger grp $\sim G$ (usually treatment grp) $y_1, \dots, y_m$ (independent samples)

Ex
$x_1 = 1000$ $y_1 = 1400$
$x_2 = 1380$ $y_2 = 1600$
$x_3 = 1200$ $y_3 = 1180$
$y_4 = 1220$

Sort $\{x_1, \dots, x_n, y_1, \dots, y_m\}$ from smallest to largest

1000, 1180, 1200, 1220, 1380, 1400, 1600
1 2 3 4 5 6 7

T_X, T_Y are rank sums of X, Y.

ex $T_X = 9$, $T_Y = 19$

Null $F = G$ (compare with $\mu_X = \mu_Y$ in t-test)
Alt $F \neq G$

If Null true, T_X or T_Y shouldn't be too big or small

Ex We test whether a diuretic works, for $\alpha = .1$, one sided test.

Two grps X = control Y = treatment

2 indep samples x_1, x_2, x_3 y_1, y_2, y_3, y_4

Null the treatment doesn't work so the control and treatment populations are the same

alt treatment work (so you pee more).

We rank daily urine production from smallest = 1 to largest = 7

We find rank sum of X (smaller grp)

T_x = 9 ← rank sum of grp X.

How extreme is this?

Find sampling distribution of T_x.

There are $\binom{7}{3} = 35$ equally likely rank sums of X.

We see that $T_x = 9$ has 1

one side p-value of $\frac{7}{35} = .2$ ← see distribution of T_x on next page

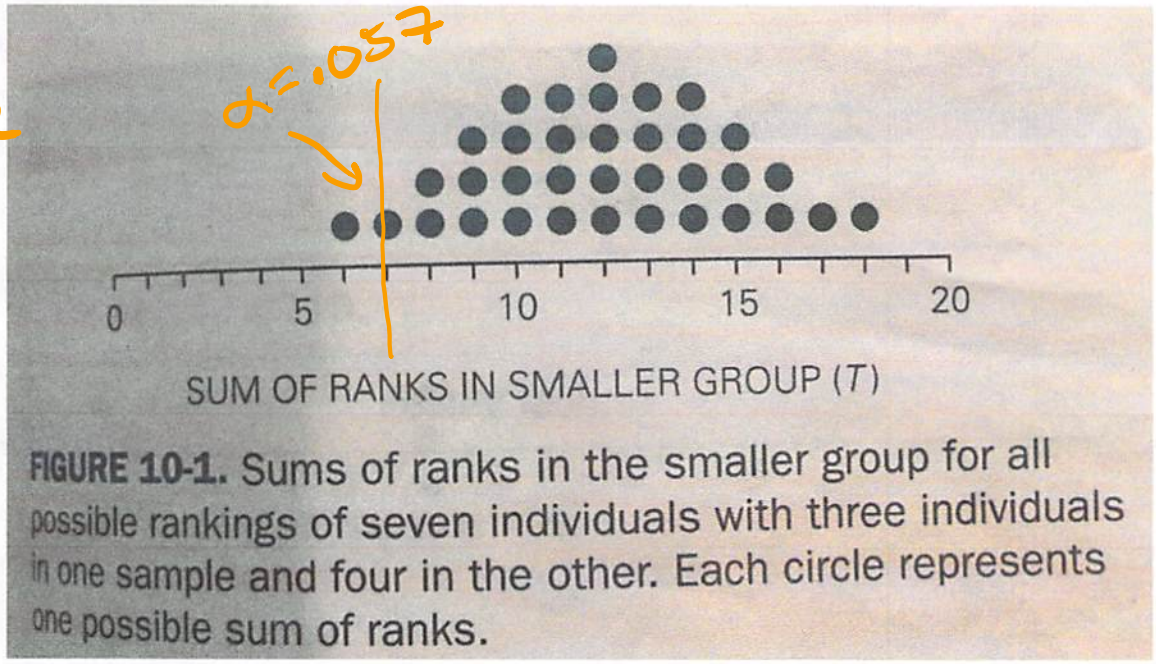
The cutoff for $\alpha = .1$ is $T_x = 7$ ← see table on next next page

and 9 7 so accept the null.

Example of Wilcoxin Rank Sum test from Glantz's Primer of Biostatistics

Placebo (Control) = X		Drug (Treatment) = Y	
Daily Urine Production (mL/d)	Rank*	Daily Urine Production (mL/d)	Rank*
1000	1	1400	6
1380	5	1600	7
1200	3	1180	2
		1220	4
	$T = 9$		

*1 = smallest; 7 = largest.



$\frac{7}{35} = \frac{1}{5} = .2$

$\alpha = .1$

$\frac{2}{35} = .057$

$\frac{4}{35} = .11$

FIGURE 10-1. Sums of ranks in the smaller group for all possible rankings of seven individuals with three individuals in one sample and four in the other. Each circle represents one possible sum of ranks.

P A21

TABLE 8 Critical Values of Smaller Rank Sum for the Wilcoxon Mann-Whitney Test

n_2	α for Two-Sided Test	α for One-Sided Test	n_1 (Smaller Sample) 1	2	3	4	5	6	7	8	9	10	11	12
3	.20	.10		3	7									
	.10	.05			6									
	.05	.025												
	.01	.005												
4	.20	.10		3	7	13								
	.10	.05			6	11								
	.05	.025				10								
	.01	.005												
5	.20	.10		4	8	14	20							
	.10	.05		3	7	12	19							
	.05	.025			6	11	17							
	.01	.005					15							
6	.20	.10		4	9	15	22	30						
	.10	.05		3	8	13	20	28						
	.05	.025			7	12	18	26						
	.01	.005				10	16	23						
7	.20	.10		4	10	16	23	32	41					
	.10	.05		3	8	14	21	29	39					
	.05	.025			7	13	20	27	36					
	.01	.005				10	16	24	32					
8	.20	.10		5	11	17	25	34	44	55				
	.10	.05		4	9	15	23	31	41	51				
	.05	.025		3	8	14	21	29	38	49				
	.01	.005				11	17	25	34	43				
9	.20	.10	1	5	11	19	27	36	46	58	70			
	.10	.05		4	*10	16	24	33	43	54	66			
	.05	.025		3	8	14	22	31	40	51	62			
	.01	.005			6	11	18	26	35	45	56			

(continued)

this is cutoff for one sided test w $\alpha = .1$ in our example.

We could have tried to show that T_y is extremely large instead of T_x being extremely small. T_x and T_y are related.

$$T_x + T_y = 1+2+3+\cdots+7 = \frac{7 \cdot 8}{2}$$

We used T_x in this problem since it is the smaller rank sum and the table above gives the critical value for the smaller rank sum.

Convention!
Two populations $X \sim F$ X = control $Y \sim G$ Y = treatment
Two indep samples $x_1, \ldots, x_n$ $y_1, \ldots, y_m$

} You book discusses T_y not T_x.

Wilkoxin tried to estimate $\pi = P(X < Y)$ and in so doing rediscovered the Mann-Whitney test.

It is useful to show how he did it because many people dont use T_x, or T_y to test H_0 but rather the Wilkoxin T.S.

To approximate $\pi = P(X < Y)$, Wilkoxin used $\hat{\pi} = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} I(x_i < y_j)$,

U_y called Wilkoxin T.S.
(called W in R).

The $I(X_i < Y_j)$ for large $n+m$ are almost independent so CLT works well and U_Y is approx normal,

We will show that

$X_1 - Y_2, X_2 - Y_2$ are dependent RVs.
but $X_1 - Y_2, X_2 - Y_2$ are indep.

$$U_Y = T_Y - \frac{n(n+1)}{2}$$

So,

$$T_Y = U_Y + \frac{n(n+1)}{2} \overset{\text{approx}}{\sim} N\left(\frac{n(n+m+1)}{2}, \frac{mn(m+n+1)}{12}\right)$$

by thm A above

How does U_Y relate to T_Y?

let $X_{(1)}, \ldots, X_{(n)}$ be the order statistic of $X_1, \ldots, X_n$

ex $x_1 = 2, x_2 = -1, x_3 = 0$

$x_{(1)} = -1, x_{(2)} = 0, x_{(3)} = 2$

Note $\sum_i \sum_j I(X_i < Y_j)$ is the

Sum of indicators of all comparisons of X and Y so

$$\sum_i \sum_j I(X_i < Y_j) = \sum_i \sum_j I(X_{(i)} < Y_{(j)})$$

Let R_{Y_1} = the rank of $Y_{(1)}$ in the combined sample $= \#(X_i' s < Y_{(1)}) + 1$

R_{Y_2} = the rank of $Y_{(2)}$ in the combined sample $= \#(X_i' s < Y_{(2)}) + 2$

$\vdots$

etc

We can write

$$U_Y = \sum_{i=1}^{n} \sum_{j=1}^{m} I(X_{(i)} < Y_{(j)})$$

$$= (\text{number of } X_i' s < Y_{(1)}) \quad (= R_{Y_1} - 1)$$

$$+ (\text{number of } X_i' s < Y_{(2)}) \quad (= R_{Y_2} - 2) + \cdots$$

$$+ (\text{number of } X_i' s < Y_{(m)}) \quad (= R_{Y_m} - m)$$

$$U_Y = \sum_{i=1}^{m} R_{Y_i} - (1+2+\cdots+m)$$

$\sum_{i=1}^{m} R_{Y_i} = T_Y$ = Sum of rank of all $Y_{(i)}$ = Sum of rank of Y_i

$(1+2+\cdots+m) = \frac{m(m+1)}{2}$

$$\Rightarrow \boxed{U_Y = T_Y - \frac{m(m+1)}{2}}$$

Fact Thm A p438

If $F = G$ (null)

$$E(T_Y) = \frac{m(m+n+1)}{2}$$

$$\text{Var}(T_Y) = \frac{mn(m+n+1)}{12}$$

pf/ easy see book.

U_Y normal $\Rightarrow T_Y$ normal ☺

also $T_Y \overset{\text{approx}}{\sim} N\left(\frac{m(m+n+1)}{2}, \frac{mn(m+n+1)}{12}\right)$

$$\boxed{U_Y \sim N\left(\frac{mn}{2}, \frac{mn(m+n+1)}{12}\right)}$$

called Wilcoxon W T.S. in R.

Stat 135 lec 28

Last time

Sec 11.2.3 The Mann Whitney Test. (nonparametric test)

Notation

X small grp $\sim F$ (usually control) $x_1, \dots, x_n$

Y larger grp $\sim G$ (usually treatment grp) $y_1, \dots, y_m$

(independent samples)

T_X, T_Y are rank sums of X, Y.

Null $F = G$

Alt $F \neq G$

If null true expect T_X not to be extreme.

Table 8 p A21 gives critical values of T_X.

Have $T_X + T_Y = 1 + 2 + 3 + \dots + (m+n) = \frac{(m+n)(m+n+1)}{2}$

For large $m+n$, assuming the Null is true,

$$T_Y \overset{\text{approx}}{\sim} N\left(\frac{m(m+n+1)}{2}, \frac{mn(m+n+1)}{12}\right)$$

The Wilcoxon T.S is

$$U_Y = \sum_{i=1}^{n}\sum_{j=1}^{m} I(x_i < y_j) \overset{\text{assuming the null is true}}{\sim} N\left(\frac{mn}{2}, \frac{mn(m+n+1)}{12}\right)$$

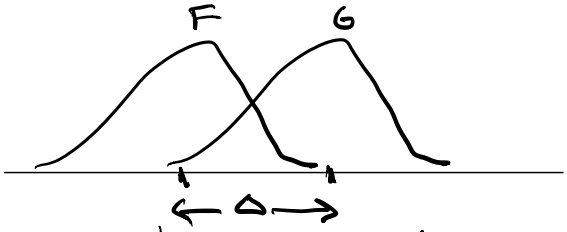
these are related by $U_Y = T_Y - \frac{m(m+1)}{2}$

Wilcoxon T.S. (called W) in R.

Today

(1) Sec 11.2.3 Wilcoxon Rank Sum test in R

Given 2 indep samples $x_1, \dots, x_n$ and $y_1, \dots, y_m$
wilcox.test( ) finds the T.S. W (this is U_y)
If willing to assume F and G are shifts of one another (i.e $G(x) = F(x-\Delta)$)

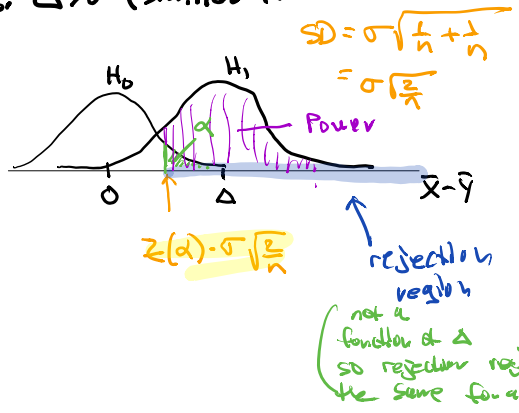


then wilcox.test provides a 95% CI of what Δ is.

Note t test is uniformly most powerful for $\Delta \neq 0$

$H_0: \mu_x - \mu_y = 0$
$H_1: \mu_x - \mu_y = \Delta \neq 0$

Picture for $\Delta > 0$ (similar for $\Delta < 0$).



So if assumptions for t-test are satisfied, it is more powerful than non-parametric test

(2) Sec 11.3.2 The Signed Rank Test

① Sec 11.2.3 Wilcoxon Rank Sum test in R
In-class exercise.

Unpaired 2 sample tests

✓ a: (data cleaning)

This problem is #21b in Rice chapter 11 exercises.

A study was done to compare the performance of engine bearings made of different compounds. Ten bearings of each type were tested. The table below gives the times until failure.

Discuss with your neighbor the code below:

```
library(tidyr)
#time to failure for two types of engine bearings
type1 <- c(3.03,5.53,5.60,9.30,9.92,12.51, 12.95,15.21,16.04,16.84)
type2 <- c(3.19, 4.26, 4.47, 4.53, 4.67, 4.69, 12.78, 6.79, 9.37, 12.75)
df <- data.frame(type1,type2)
head(df)
```

	type1 <dbl>	type2 <dbl>
1	3.03	3.19
2	5.53	4.26
3	5.60	4.47
4	9.30	4.53
5	9.92	4.67
6	12.51	4.69

6 rows

```
#we reformat the data frame to tidy format for easier analysis
df_tidy<- df %>% gather(key=type, value=time, `type1`, `type2`)
head(df_tidy)
```

	type <chr>	time <dbl>
1	type1	3.03
2	type1	5.53
3	type1	5.60
4	type1	9.30
5	type1	9.92
6	type1	12.51

6 rows

✓ b: nonparametric Mann-Whitney test (Wilcoxin rank sum test)

We test the hypothesis with a nonparametric method (Wilcoxin rank sum test).

H_0: type 1 distribution is the same as type 2 distribution

H_1: type 1 distribution is different from type 2 distribution.

We compute the rank sum for the smaller size group (in this case both groups are the same) and the smaller rank sum.

```
df_tidy <- df_tidy %>% mutate(ranks=rank(time))
df_tidy
```

type <chr>	time <dbl>	ranks <dbl>
type1	3.03	1
type1	5.53	8
type1	5.60	9
type1	9.30	11
type1	9.92	13
type1	12.51	14
type1	12.95	17
type1	15.21	18
type1	16.04	19
type1	16.84	20

1-10 of 20 rows Previous 1 2 Next

```
sum_rank_df <- df_tidy %>% group_by(type) %>% summarize(sum_rank=sum(ranks))
sum_rank_df
```

type <chr>	sum_rank <dbl>
type1	130
type2	80

2 rows

To test our hypothesis exactly, we use look up table with two sided alpha=.05, $n_1 = n_2 = 10$ on page A22 in Rice. We see that the cutoff is 78. We got 80 so we accept the null.

Lets compare this with the normal approximation of the rank sum. With $m = n = 10$ we have

$$T \sim N(m(m+n+1)/2, mn(m+n+1)/12) = N(105, 175).$$

We compute the two sided p-value of rank sum T=80.

```
pval=2*(1-pnorm(abs((80-105))/sqrt(175)))
pval
```

```
## [1] 0.05878172
```

The p-value is slightly larger than the $\alpha = .05$ so we accept the null.

c: t-test for unpaired data

This is 11.6.21a where we do a t-test to compare the means of two normal distributions.

```
#check whether we can assume equal variance
var(type1)
```

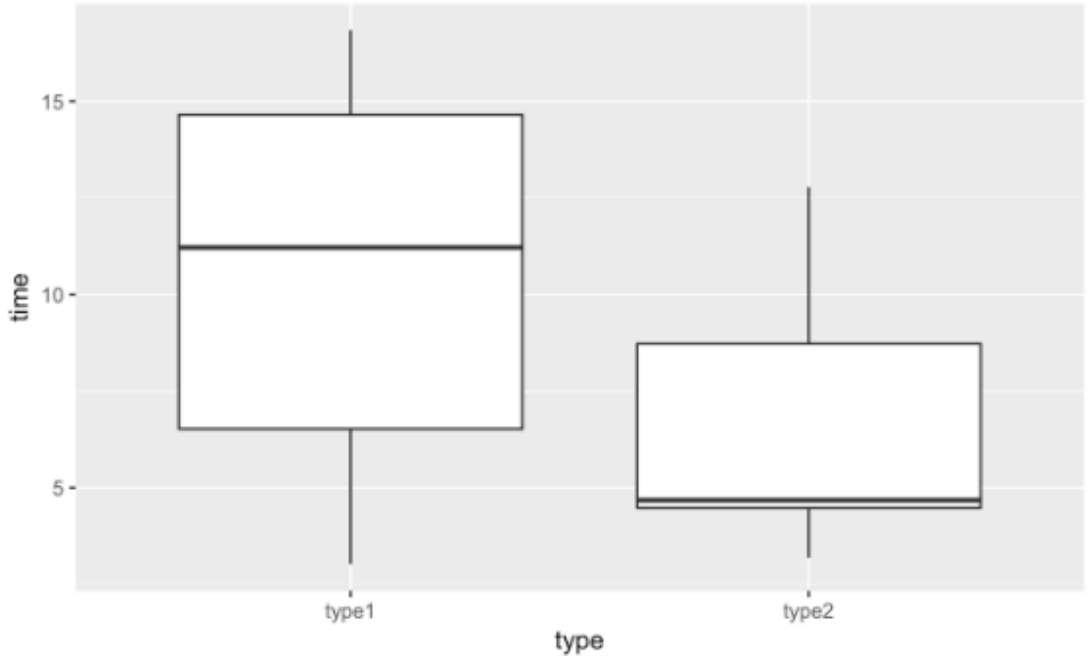
```
## [1] 23.22551
```

```
var(type2)
```

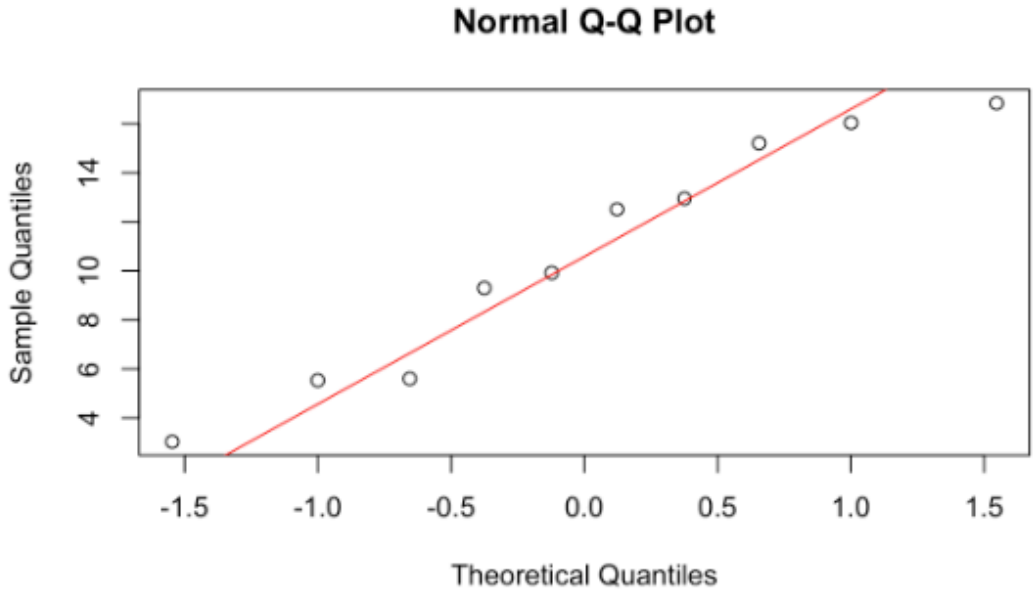
```
## [1] 12.97749
```

variances appear to be different

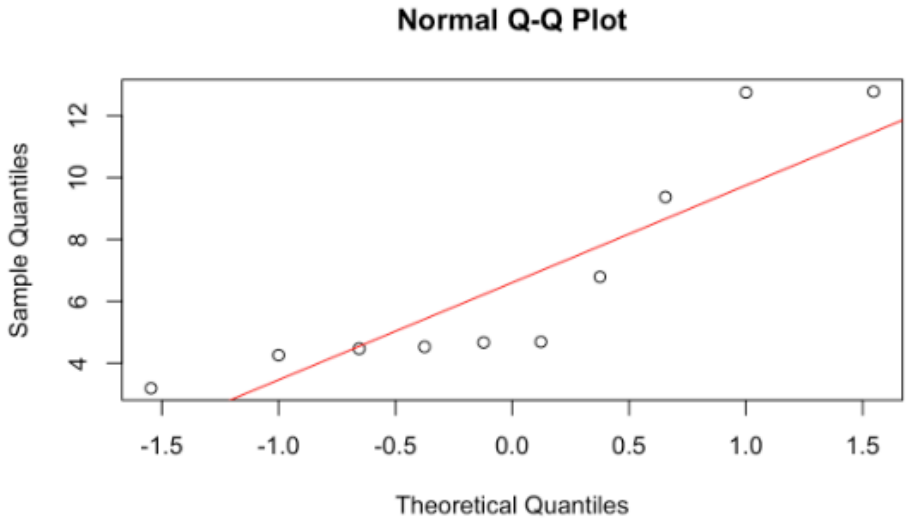
```
df_tidy %>% ggplot(aes(x=type,y=time)) + geom_boxplot()
```




```
qqnorm(type1); qqline(type1,col=2)
```



```
qqnorm(type2); qqline(type2,col=2)
```




```
#default is paired=FALSE, var.equal=FALSE, conf.level=0.95
t.test(alternative="two.sided",type1,type2)
```

```
##
##  Welch Two Sample t-test
##
## data:  type1 and type2
## t = 2.0723, df = 16.665, p-value = 0.05408
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.07752643  7.96352643
## sample estimates:
## mean of x mean of y
##    10.693     6.750
```

Note that the p-value for the parametric t.test is close to the p-value for the nonparametric Mann-Whitney test. This indicates that the assumptions of the t-test aren't badly violated.

We know that the t-test is uniformly most powerful when the assumptions of the t-test are met (i.e. the rejection region of the t-test isn't a function of the value of the alternative). Under these conditions it is known that the Mann Whitney test is almost as powerful (95% as powerful). When the assumptions of the t-test are not satisfied the results of the t-test can be misleading and we shouldn't really talk about the power of the test. In this case the Mann Whitney test is prefered.

We can also do a Mann-Whitney test with the built in Wilcoxin.test() function in R. Unfortunately this test computes a test statistic, $W = T - m(m+1)/2 = 80 - 55 = 25$. W is the number of pairs of type1 and type 2 where type2 is less than type1. Out of 100 we would expect 50 if the null is true and we got 25. This isn't too extreme.

```
wilcox.test(type2, type1, alternative="two.sided", exact=FALSE,conf.int=.95, correct=FALSE)
```

```
##
##  Wilcoxon rank sum test
##
## data:  type2 and type1
## W = 25, p-value = 0.05878
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -8.42001523  0.07004542
## sample estimates:
## difference in location
##              -4.27779
```

We accept the null that the two distributions are the same.

Which test is better in this case, t-test or nonparametric test? In my opinion you have a lot to gain for using the Mann-Whitney test and not much to lose.

Sec 11.3.2 Non Parametric Wilcoxin Signed Rank Test.

For Paired data.

B	A	D_i (A-B)	$\lvert D_i \rvert$	Rank	Signed Rank
25	27	2	2	2	2
29	25	-4	4	3	-3
60	59	-1	1	1	-1
27	37	10	10	4	4

$W_+ = 2+4=6$ ← sum of pos signed ranks.

Called V in R

H_0: Distribution of D_i is symmetric around 0

H_1: " " " is not symmetric around 0

Thm Under the null that D_i are symmetrically distributed around 0

$$W_+ \sim N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$$

Pf/ see book.

Here it is in R:

Paired 2 sample tests

This is problem 48 in Rice. Discuss the following code with your neighbor.

The amount of urinary protein before and after a certain treatment is given in the table below:

```
before <- c(24.6,17,16,10.4,8.2,7.9,8.2,7.9,5.8,5.4,5.1,4.7)
after <- c(10.1,5.7,5.6,3.4,6.5,0.7,6.5,0.7,6.1,4.7,2.0,2.9)
diff=before-after
abs_diff=abs(diff)
rank <- rank(abs_diff)
signed_rank <- rank*(diff/abs_diff)
df <- data.frame(before,after,diff, abs_diff,rank, signed_rank)
df
```

before <dbl>	after <dbl>	diff <dbl>	abs_diff <dbl>	rank <dbl>	signed_rank <dbl>
24.6	10.1	14.5	14.5	12.0	12.0
17.0	5.7	11.3	11.3	11.0	11.0
16.0	5.6	10.4	10.4	10.0	10.0
10.4	3.4	7.0	7.0	7.0	7.0
8.2	6.5	1.7	1.7	3.5	3.5
7.9	0.7	7.2	7.2	8.5	8.5
8.2	6.5	1.7	1.7	3.5	3.5
7.9	0.7	7.2	7.2	8.5	8.5
5.8	6.1	-0.3	0.3	1.0	-1.0
5.4	4.7	0.7	0.7	2.0	2.0

1-10 of 12 rows Previous 1 2 Next

null and alternative:

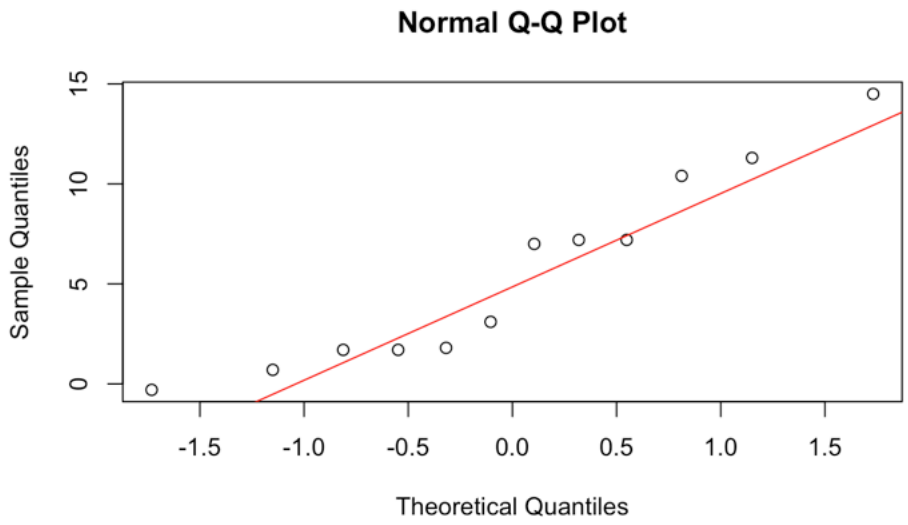
H_0 : the distribution of D_i is symmetric around zero

H_1 : the distribution of D_i isn't symmetric around zero

✓ a) t-test

check assumption that D_i are iid normal:

```
qqnorm(diff);qqline(diff,col=2)
```



If assumptions that the differences looks normal we perform a t-test using command t.test and make conclusion:

```
#step2
t.test(before,after,paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  before and after
## t = 4.0012, df = 11, p-value = 0.002082
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.485818 8.564182
## sample estimates:
## mean of the differences
##                   5.525
```

b) Wilcoxin Signed Rank test

Use the command `wilcox.test()` to perform a Wilcoxin Signed Rank test. Use console below for your answer

Code | Start Over | Solution | Run Code

```
1 wilcox.test(before,after,paired = TRUE,correct=FALSE)
2
3
```

before <dbl>	after <dbl>	diff <dbl>	abs_diff <dbl>	rank <dbl>	signed_rank <dbl>
24.6	10.1	14.5	14.5	12.0	12.0
17.0	5.7	11.3	11.3	11.0	11.0
16.0	5.6	10.4	10.4	10.0	10.0
10.4	3.4	7.0	7.0	7.0	7.0
8.2	6.5	1.7	1.7	3.5	3.5
7.9	0.7	7.2	7.2	8.5	8.5
8.2	6.5	1.7	1.7	3.5	3.5
7.9	0.7	7.2	7.2	8.5	8.5
5.8	6.1	-0.3	0.3	1.0	-1.0
5.4	4.7	0.7	0.7	2.0	2.0

1-10 of 12 rows Previous 1 2 Next

```
Warning in wilcox.test.default(before, after, paired = TRUE, correct =
FALSE): cannot compute exact p-value with ties
```

```
    Wilcoxon signed rank test

data:  before and after
V = 77, p-value = 0.002852
alternative hypothesis: true location shift is not equal to 0
```

Note that the test statistic is called `V` in R. Which test is better in this case, t-test or nonparametric test?

If assumptions are met, t-test is a uniformly most powerful test. The Wilcoxin Signed Rank Test is around 95% as powerful.

ex (See 11.6.266) — HW #8

Let $X_1 \sim N(0,1)$, $Y_1, Y_2 \sim N(1,1)$ } Indep samples

Determine the variance of the rank sum of the X's.

i.e. $Var(T_x)$.

$$U_y = T_y - \frac{n(n+1)}{2}$$

$$T_x + T_y = \frac{(m+n)(m+n+1)}{2}$$

So $Var(T_x) = Var(U_y)$

So we find $Var(U_y)$.

$$U_y = \sum_i \sum_j I(x_i < y_j)$$

$x_1 - y_1$, $x_1 - y_2$ are dependent

$$Var(U_y) = Var\left(I(x_1<y_1) + I(x_1<y_2)\right)$$

$$= Var\left(I(x_1<y_1)\right) + Var\left(I(x_1<y_2)\right) + 2Cov\left(I(x_1<y_1), I(x_1<y_2)\right)$$

$$Cov\left(I(x_1<y_1), I(x_1<y_2)\right) = E\left(I(x_1<y_1)I(x_1<y_2)\right) - E\left(I(x_1<y_1)\right)E\left(I(x_1<y_2)\right)$$

$$E\left(I(x_1<y_1)I(x_1<y_2)\right) = P(x_1 < y_1 \text{ and } x_1 < y_2)$$

↖ hard to find analytically so lets approx it by simulation in R.

R simulation

```
B=10000

fun <- function(){
 a <- rnorm(1)
 b <- rnorm(2,mean = 1)
 p <- prod(a<b)
}

vec <- replicate(B,fun())
mean(vec)
```

x_1 → a

y_1, y_2 → b

1 or 0 → p $= I(x_1<y_1)\cdot I(x_1<y_2)$

You will find that $P(x_1 < y_1 \text{ and } x_1 < y_2) = .634$

and $Cov\left(I(x_1<y_1), I(x_1<y_2)\right) = .634 - (.76)^2 = .056$

$P(x_1 < y_1) = .76$

$$Var(U_y) = 2(.76)(.24) - 2(.056)$$

Stat 135 Lec 29

Last time

Sec 11.2 and 11.3 nonparametric tests for comparing the distribution of 2 continuous populations

Today

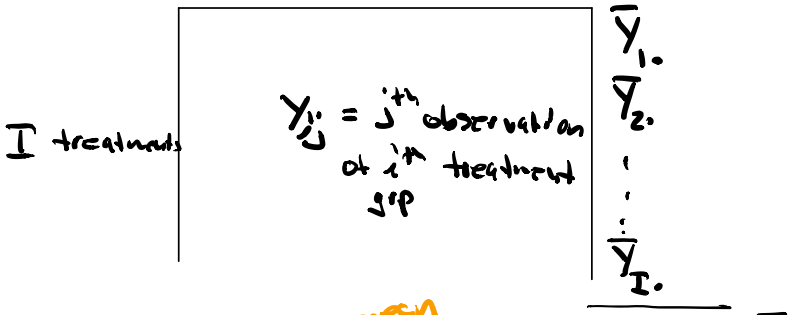
Sec 12.2 Analysis of Variance (ANOVA)

parametric test (Normal theory F-test) comparing the means of 2 or more continuous populations.

Sec 12.2 ANOVA (Analysis of Variance)

One-way layout is an experimental design in which independent measurements are made under each of several treatments

Picture



model

overall mean, differential effect of i^{th} treatment

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

$$\sum_{i=1}^{I} \alpha_i = 0$$

$$Y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$$

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_I = 0 \qquad E(Y_{ij}) = \mu$

H_1: not all α_i are zero.

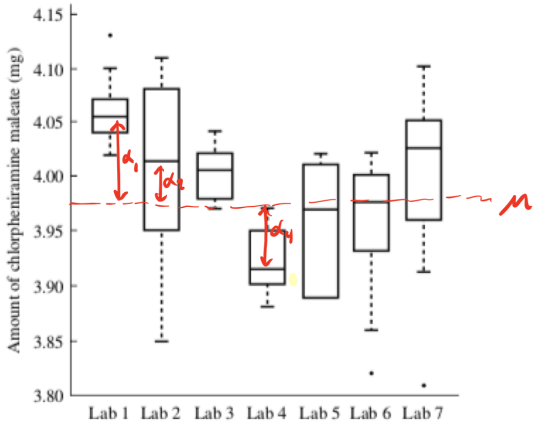


FIGURE 12.1 Boxplots of determinations of amounts of chlorpheniramir in tablets by seven laboratories.

A nova

Test will track differential effect α_i is zero.

ANOVA is based on decomposition of the total variance (SST) (sum of squares total) of your data intra group variance (SSW) (sum of squares within) and the inter-group variance (SSB) (sum of squares between).

$$SST = SSW + SSB$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y})^2 \quad \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_{i\cdot})^2 \quad \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{Y}_{i\cdot}-\bar{Y})^2$$

The algebraic proof is on p408.

SSW and SSB are independent RVs.

to see this:

Facts from Chap 6

$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

① $\bar{X}$ and $X_j - \bar{X}$ are indep RVs (thm A p.195)

② $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ (thm B p.197)

$$SSB = \sum_{i=1}^{I}\sum_{j=1}^{J} (\bar{Y}_{i\cdot} - \bar{Y})^2$$ is a function of $\bar{Y}_{i\cdot}$ since $\bar{Y} = \frac{1}{I}\sum_{i=1}^{I} \bar{Y}_{i\cdot}$

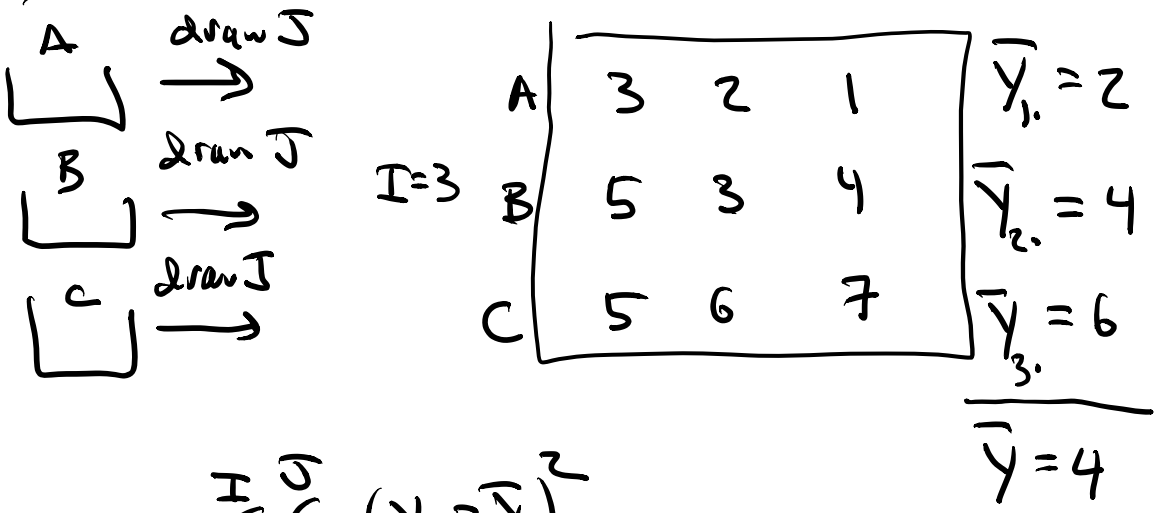
this is $\bar{X}$ above in ①

$$SSW = \sum_{i=1}^{I}\sum_{j=1}^{J} (Y_{ij} - \bar{Y}_{i\cdot})^2$$ is a function of $Y_{ij} - \bar{Y}_{i\cdot}$

this is $X_j - \bar{X}$ above in ①

$\Rightarrow$ SSW and SSB are indep RVs.

ex $I=3, J=3$



$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y})^2$$

$$= (3-4)^2+(2-4)^2+(1-4)^2$$
$$+ (5-4)^2+(3-4)^2+(4-4)^2$$
$$+ (5-4)^2+(6-4)^2+(7-4)^2$$
$$= \boxed{30}$$

this is where we assume the null.

Under the null we have $I \cdot J$ draws from a single box,

$$\frac{SST}{IJ-1} = S^2 \Rightarrow \frac{(IJ-1)S^2}{\sigma^2} \sim \chi^2_{IJ-1}$$

we say SST has $IJ-1$ d.f assuming the null

SSW $\qquad J = 3$

$I=3$	A	3	2	1	$\bar{Y}_{1\cdot} = 2$
	B	5	3	4	$\bar{Y}_{2\cdot} = 4$
	C	5	6	7	$\bar{Y}_{3\cdot} = 6$

$\bar{Y} = 4$

$$SSW = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{i\cdot})^2 = \underbrace{(3-2)^2 + (2-2)^2 + (1-2)^2}_{=(J-1)S_A^2} + \underbrace{(5-4)^2 + (3-4)^2 + (4-4)^2}_{=(J-1)S_B^2} + \underbrace{(5-6)^2 + (3-6)^2 + (7-6)^2}_{=(J-1)S_C^2} = 6$$

$$\leftarrow (J-1)S_A^2 + (J-1)S_B^2 + (J-1)S_C^2$$

$$\frac{SSW}{\sigma^2} = I(J-1)\frac{(J-1)S_A^2 + (J-1)S_B^2 + (J-1)S_C^2}{I(J-1)} = S^2_{pooled}$$

$$\frac{\qquad\qquad}{\sigma^2} \sim \chi^2_{I(J-1)}$$

We say SSW has $I(J-1)$ d.f.

SSB

$J = 3$

$I=3$	A	3	2	1	$\bar{Y}_{1\cdot} = 2$
	B	5	3	4	$\bar{Y}_{2\cdot} = 4$
	C	5	6	7	$\bar{Y}_{3\cdot} = 6$
					$\bar{Y} = 4$

$$SSB = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\bar{Y}_{i\cdot} - \bar{Y}\right)^2$$

$$= (2-4)^2 + (2-4)^2 + (2-4)^2$$
$$+ (4-4)^2 + (4-4)^2 + (4-4)^2$$
$$+ (6-4)^2 + (6-4)^2 + (6-4)^2 = 24$$

Notice

$$SST = SSW + SSB$$
$$11 \qquad 11 \qquad 11$$
$$30 \qquad 6 \qquad 24$$

✓

Facts

1) $SST = SSW + SSB$

2) SSW and SSB are indep

3) chi square subtraction property

If $X = X_1 + X_2$ with X_1, X_2 indep

and $X \sim \chi^2_{n_1+n_2}$

$X_1 \sim \chi^2_{n_1}$

then $X_2 \sim \chi^2_{n_2}$.

have $n_1 =$ df of $SSW = I(J-1)$

$n_1+n_2 =$ df of $SST = IJ-1$

$\Rightarrow n_2 =$ df of $SSB = IJ-1 - I(J-1)$

$= \boxed{I-1}$

We say SSB has $I-1$ d.f. assuming null.

(we need null assumption since we use fact that SST has $IJ-1$ d.f under null.).

F distribution — Chap 6 p194

Recall χ^2, t, F are distributions based on the normal distribution.

Def^n Let U and V be indep χ^2 RV w/ a and b degrees of freedom, the distribution

$$F_{a,b} = \frac{U/a}{V/b}$$ is called the F distribution

F test (assuming the null)

$$F_{I-1,\, I(J-1)} = \frac{SSB/\sigma^2}{I-1} \Big/ \frac{SSW/\sigma^2}{I(J-1)}$$

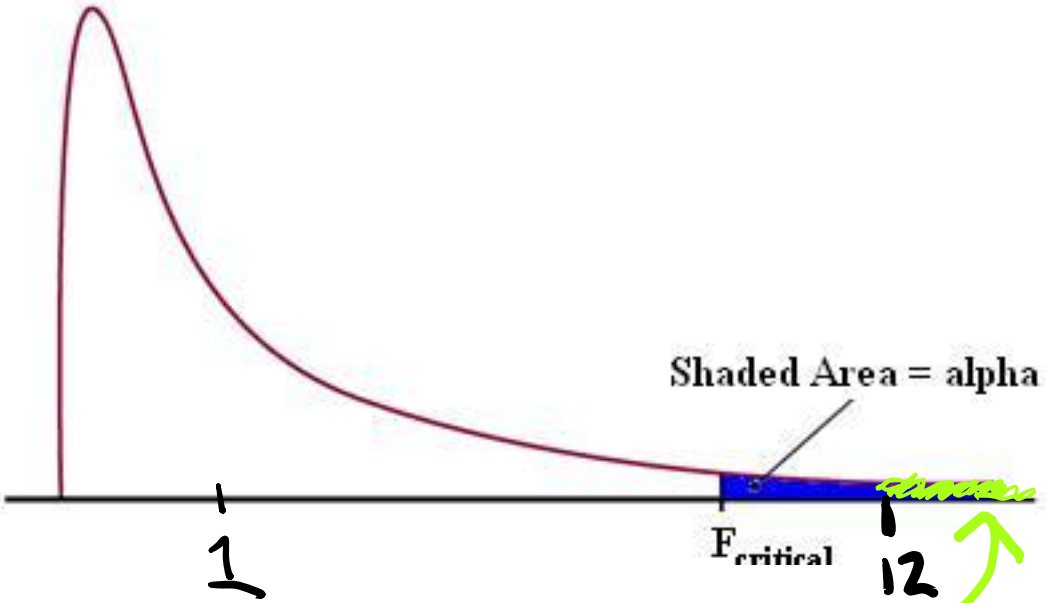
$I-1$: between d.f. ; $I(J-1)$: within d.f.

$$= \frac{SSB/I-1}{SSW/I(J-1)}$$

For our example

$$F_{2,6} = \frac{24/2}{6/6} = 12$$

$F_{2,6}$ distribution.



In R

1 - Pf(12, 2, 6) = .008

(2 ↑ df between; 6 ↑ df within)

So we reject null that all treatments have same mean.

Stat 135 lec 30

Last time

Sec 12.2 One way ANOVA with diffent size grps $J_1, J_2, \dots, J_I$

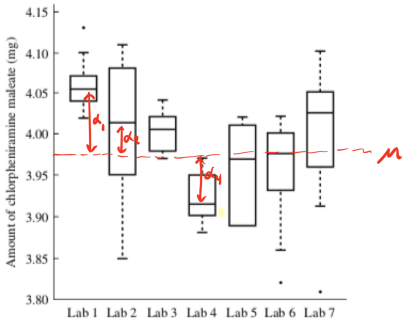


FIGURE 12.1 Boxplots of determinations of amounts of chlorpheniramir in tablets by seven laboratories.

J

I

model

$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma^2)$

μ: overall mean; α_i: differential effect of i^{th} treatment

A nova

$$SST = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \bar{Y})^2 \quad \text{and} \quad \frac{SST}{\sigma^2} \sim \chi^2_{IJ-1}$$

$IJ - 1 = \sum_{i=1}^{I} J_i - 1$

assuming null that all $\alpha_i = 0$

similarly

$$SSW = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{i\cdot})^2 \quad \text{and} \quad \frac{SST}{\sigma^2} \sim \chi^2_{I(J-1)}$$

$I(J-1) = \sum_{i=1}^{I} (J_i - 1)$

$$SSB = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (\bar{Y}_{i\cdot} - \bar{Y})^2 \quad \text{and} \quad \frac{SSB}{\sigma^2} \sim \chi^2_{I-1}$$

assuming the null

F test

$$F_{\left(I-1, \sum_{i=1}^{I}(J_i-1)\right)} = \frac{SSB/I-1}{SSW/\sum_{i=1}^{I}(J_i-1)}$$

assuming the null

Today

0. Assumptions
1. Shortcut Bonferroni
2. Kruskal Wallis (KW) ~~test~~ (non parametric)
3. KW test in R.

⓪ Assumptions

The normal theory F-test has several assumptions

1) our data Y_{ij} is normal
2) the variance of each treatment group is the same σ^2
3) All observations are independent.

If we dont satisfy these use non parametric test (Kruskal Wallis test).

① Shortcut Bonferroni test.

Bonferroni test

Suppose the F test rejects the null (i.e at least one pair of means are different)

If we do $\binom{I}{2}$ simultaneous pairwise t-tests to see which pairs of treatment groups have diff. means, the type I error $P_0(\phi(x)=1)$ is additive and it is quite likely that we reject the null for one of the t-tests.

Bonferroni method is to make significance level of each test $\frac{\alpha}{\binom{I}{2}}$ level of significance.

Shortcut Bonferroni method

The Bonferroni method is a lot of work and the small significance level makes it hard to reject the null for each t-test.

Before I do the Bonferroni test I do this shortcut test:

For each treatment group $H_0: \alpha_i = 0$ $(\mu_i = \mu)$
$H_1: \alpha_i \neq 0$

we compute I simultaneous $100\left(1 - \frac{\alpha/2}{I}\right)$ two sided CI for the true mean μ_i for each group.

We simply look whether one of the CI doesn't overlap with the others.

On test or HW I will accept this method.

Ex In our 3 grp example

A	3	2	1	$\bar{y}_{1\cdot} = 2$	$(J-1)S_A^2 = (3-2)^2 + (2-2)^2 + (1-2)^2 = 2$
B	5	3	4	$\bar{y}_{2\cdot} = 4$	$(J-1)S_B^2 = 2$
C	5	6	7	$\bar{y}_{3\cdot} = 6$	$(J-1)S_C^2 = 2$
				$\bar{\bar{y}} = 4$	

$F_{2,6} = .008 < .05 = \alpha$ $\Rightarrow$ reject null

So we perform a shortcut Bonferroni.

Recall

$$\frac{\bar{Y}_{i\cdot} - \mu_i}{S/\sqrt{J_i}} \sim t_{J_i - 1}$$

$\Rightarrow$ A $100\left(1 - \frac{\alpha/2}{I}\right)\%$ CI for μ_i is

$$\bar{Y}_{i\cdot} \pm t_{J_i - 1}\left(\frac{\alpha/2}{I}\right) \cdot S/\sqrt{J_i}$$

In our example we had $S_i \approx 1$.

For $\frac{\alpha}{2} = .025$ there a special Bonferroni table $t\left(\frac{\alpha/2}{I}\right)_{J_i - 1}$

← in table next page

↖ # treat.

↑ d.f

so

$$\bar{Y}_{1\cdot} \pm t_2\left(\frac{.025}{3}\right) \cdot \frac{1}{\sqrt{3}} = 2 \pm 4.41 = 7.648\left(\frac{1}{\sqrt{3}}\right)$$

$$\bar{Y}_{2\cdot} \pm t_2\left(\frac{.025}{3}\right) \cdot \frac{1}{\sqrt{3}} = 4 \pm 4.41$$

$$\bar{Y}_{3\cdot} \pm t_2\left(\frac{.025}{3}\right) \cdot \frac{1}{\sqrt{3}} = 6 \pm 4.41$$

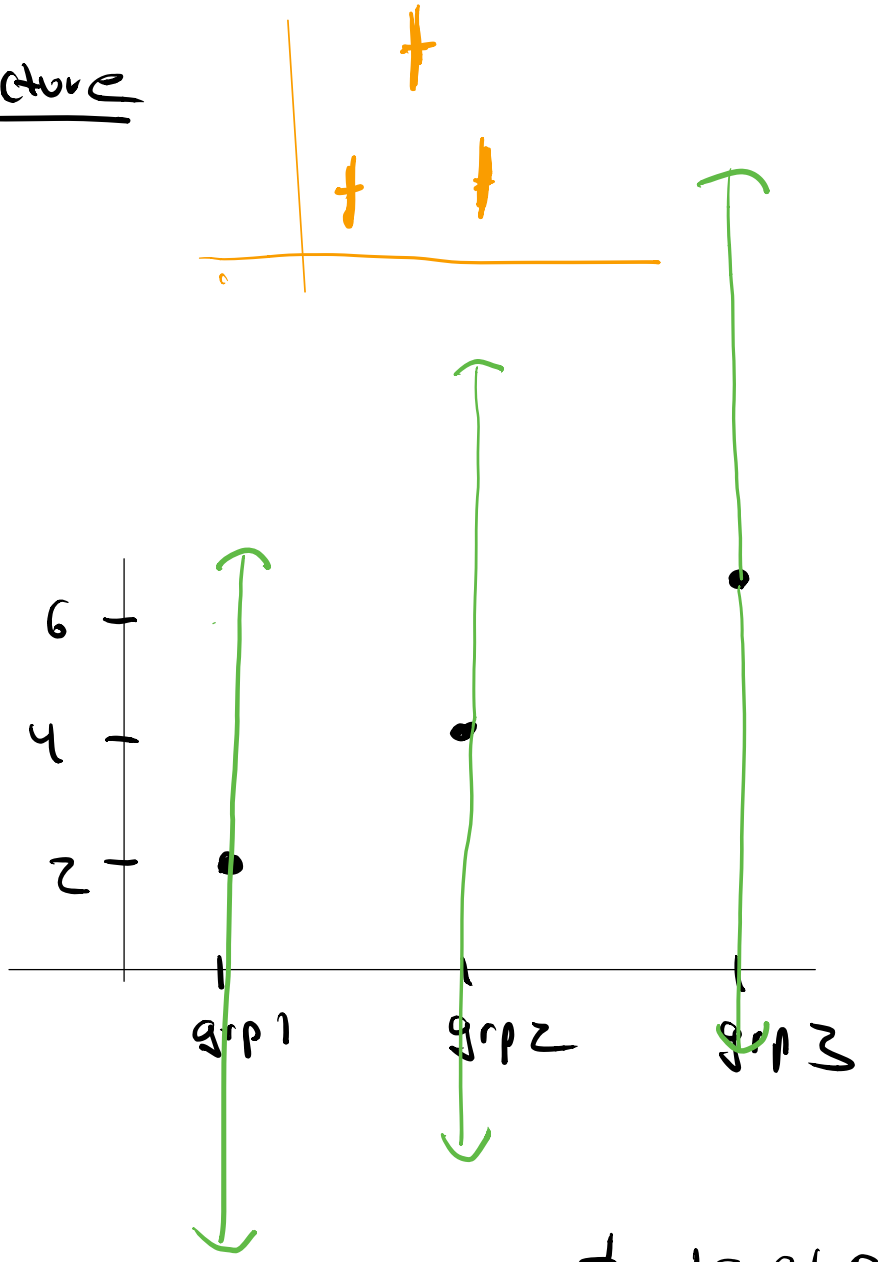
Bonferonni table (not in book)

TABLE 11 Bonferroni Multipliers for 95% Confidence Intervals

The values given in the table are $t_{df,0.025/k}$ where k is the number of tests.

df	NUMBER OF TESTS 1	2	3	4	5	6	8	10	15	20
1	12.706	25.452	38.185	50.923	63.657	76.384	101.856	127.321	190.946	254.647
2	4.303	6.205	7.648	8.860	9.925	10.885	12.590	14.089	17.275	19.963
3	3.182	4.177	4.857	5.392	5.841	6.231	6.895	7.453	8.575	9.465
4	2.776	3.495	3.961	4.315	4.604	4.851	5.261	5.598	6.254	6.758
5	2.571	3.163	3.534	3.810	4.032	4.219	4.526	4.773	5.247	5.604
6	2.447	2.969	3.287	3.521	3.707	3.863	4.115	4.317	4.698	4.981
7	2.365	2.841	3.128	3.335	3.499	3.636	3.855	4.029	4.355	4.595
8	2.306	2.752	3.016	3.206	3.355	3.479	3.677	3.833	4.122	4.334
9	2.262	2.685	2.933	3.111	3.250	3.364	3.547	3.690	3.954	4.146
10	2.228	2.634	2.870	3.038	3.169	3.277	3.448	3.581	3.827	4.005
11	2.201	2.593	2.820	2.981	3.106	3.208	3.370	3.497	3.728	3.895
12	2.179	2.560	2.779	2.934	3.055	3.153	3.308	3.428	3.649	3.807
13	2.160	2.533	2.746	2.896	3.012	3.107	3.256	3.372	3.584	3.735
14	2.145	2.510	2.718	2.864	2.977	3.069	3.214	3.326	3.529	3.675
15	2.131	2.490	2.694	2.837	2.947	3.036	3.177	3.286	3.484	3.624
16	2.120	2.473	2.673	2.813	2.921	3.008	3.146	3.252	3.444	3.581
17	2.110	2.458	2.655	2.793	2.898	2.984	3.119	3.222	3.410	3.543
18	2.101	2.445	2.639	2.775	2.878	2.963	3.095	3.197	3.380	3.510
19	2.093	2.433	2.625	2.759	2.861	2.944	3.074	3.174	3.354	3.481
20	2.086	2.423	2.613	2.744	2.845	2.927	3.055	3.153	3.331	3.455
25	2.060	2.385	2.566	2.692	2.787	2.865	2.986	3.078	3.244	3.361
30	2.042	2.360	2.536	2.657	2.750	2.825	2.941	3.030	3.189	3.300
40	2.021	2.329	2.499	2.616	2.704	2.776	2.887	2.971	3.122	3.227
50	2.009	2.311	2.477	2.591	2.678	2.747	2.855	2.937	3.083	3.184
60	2.000	2.299	2.463	2.575	2.660	2.729	2.834	2.915	3.057	3.156
70	1.994	2.291	2.453	2.564	2.648	2.715	2.820	2.899	3.039	3.137
80	1.990	2.284	2.445	2.555	2.639	2.705	2.809	2.887	3.026	3.122
100	1.984	2.276	2.435	2.544	2.626	2.692	2.793	2.871	3.007	3.102
140	1.977	2.266	2.423	2.530	2.611	2.676	2.776	2.852	2.986	3.079
1000	1.962	2.245	2.398	2.502	2.581	2.643	2.740	2.813	2.942	3.031
∞	1.960	2.241	2.394	2.498	2.576	2.638	2.734	2.807	2.935	3.023

Picture



No conclusion. The $\frac{\alpha}{I}$ level of significance the CI too big. We need J to be bigger to see non overlapping intervals.

Stat135 lecture 31

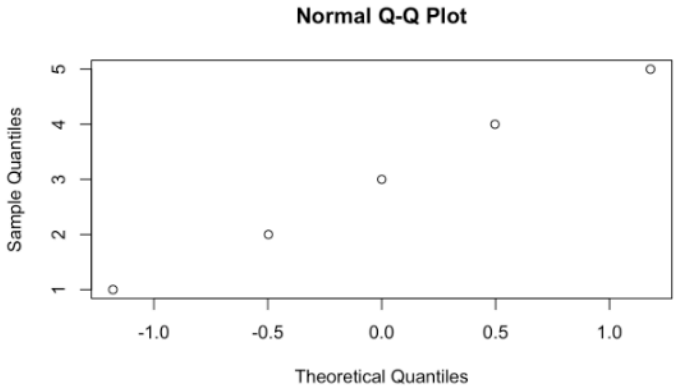
(data cleaning)

a: F-test

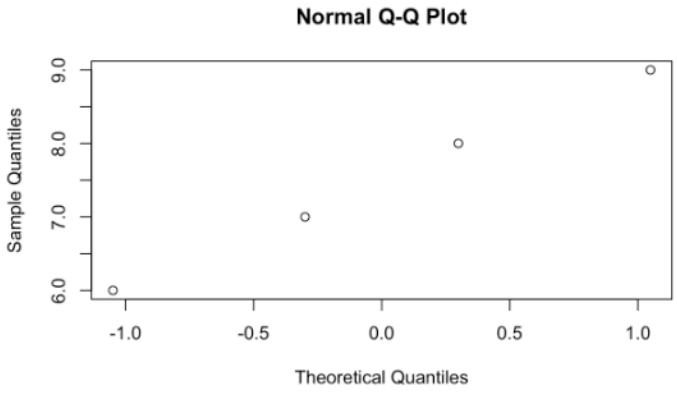
b: nonparametric Kruskal-Wallis test

Start Over

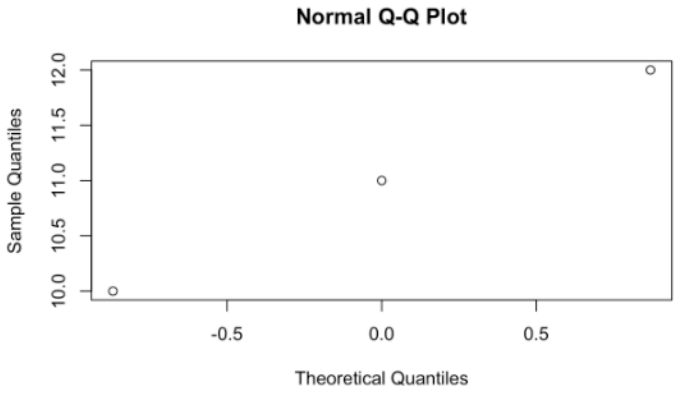
```
library(tidyr)
#time to failure for three types of engine bearings

treatment1 <- c(1,2,3,4,5)
treatment2 <- c(6,7,8,9)
treatment3 <- c(10,11,12)

#you should check that normality assumption is satisfied.
qqnorm(treatment1)
```




```
qqnorm(treatment2)
```




```
qqnorm(treatment3)
```



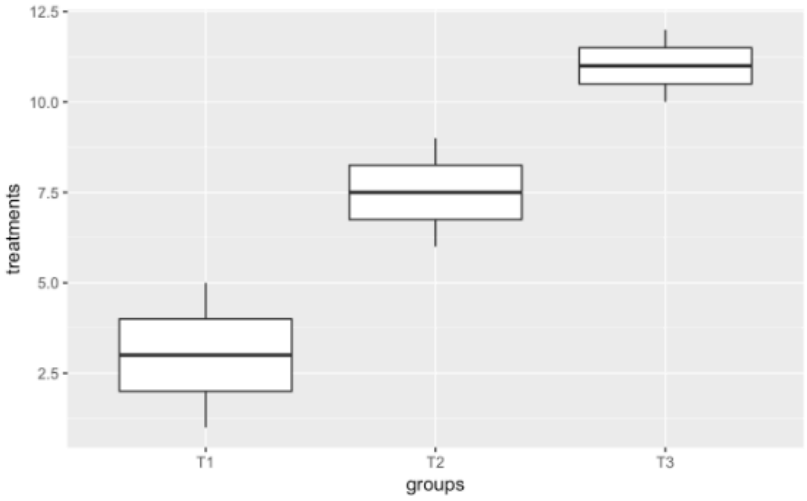
```
groups <- c(rep("T1",5),rep("T2",4),rep("T3",3))
treatments <- c(treatment1,treatment2,treatment3)
df_tidy <- data.frame(groups,treatments)
df_tidy
```

groups <fctr>	treatments <dbl>
T1	1
T1	2
T1	3
T1	4
T1	5
T2	6
T2	7
T2	8
T2	9
T3	10

1-10 of 12 rows Previous 1 2 Next

You should make boxplot to examine the means and variances

```
df_tidy %>% ggplot(aes(x=groups,y=treatments))+ geom_boxplot()
```



a: F-test

```
I=3
J=c(5,4,3)

mean_T1 <- mean(treatment1)
mean_T2 <- mean(treatment2)
mean_T3 <- mean(treatment3)
mean_T1
```

```
## [1] 3
```

```
mean_T2
```

```
## [1] 7.5
```

```
mean_T3
```

```
## [1] 11
```

```
mean_tot <- mean(df_tidy$treatments)


var_T1 <- var(treatment1)
var_T2 <- var(treatment2)
var_T3 <- var(treatment3)
var_T1
```

```
## [1] 2.5
```

```
var_T2
```

```
## [1] 1.666667
```

```
var_T3
```

```
## [1] 1
```

```
SSW <-  (J[1]-1)*var(treatment1) + (J[2]-1)*var(treatment2) + (J[3]-1)*var(treatment3)
SSW
```

```
## [1] 17
```

```
SSB <- J[1]*(mean_T1-mean_tot)^2 + J[2]*(mean_T2-mean_tot)^2 + J[3]*(mean_T3-mean_tot)^2
SSB
```

```
## [1] 126
```

```
F <- (SSB/2)/(SSW/9)
F
```

```
## [1] 33.35294
```

```
1-pf(F,2,9)
```

```
## [1] 6.886649e-05
```

Using `oneway.test()`

```
oneway.test(treatments~groups,var.equal = TRUE)  #in formula rhs: treatment values lhs: gro
up type
```

```
##
##  One-way analysis of means
##
## data:  treatments and groups
## F = 33.353, num df = 2, denom df = 9, p-value = 6.887e-05
```

Or can assume not equal variance:

```
oneway.test(treatments~groups)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  treatments and groups
## F = 34.489, num df = 2.000, denom df = 5.771, p-value = 0.000617
```

Shortcut Bonferonni method

Since we reject the null lets do the shortcut Bonferonni method (level of significance for each test is 91.7%) to see which group means are significantly different.

```
#T1
mean_T1-qt(1-.025/3,4)*sqrt(var_T1)/sqrt(5)
```

```
## [1] 0.199301
```

```
mean_T1+qt(1-.025/3,4)*sqrt(var_T1)/sqrt(5)
```

```
## [1] 5.800699
```

T_1 CI is (0.2,5.8)

```
#T2
mean_T2-qt(1-.025/3,3)*sqrt(var_T2)/sqrt(4)
```

```
## [1] 4.365041
```

```
mean_T2 +qt(1-.025/3,3)*sqrt(var_T2)/sqrt(4)
```

```
## [1] 10.63496
```

T_2 CI is (4.4, 10.6)

```
#T3
mean_T3-qt(1-.025/3,2)*sqrt(var_T3)/sqrt(3)
```

```
## [1] 6.583961
```

```
mean_T3+qt(1-.025/3,2)*sqrt(var_T3)/sqrt(3)
```

```
## [1] 15.41604
```

T_3 CI is (6.6., 15.4).

We see T1 and T3 CIs don't overlap. We conclude that the means of T_1 and T_3 are different.

lec 31

Quiz 4 Friday Nov 22 — Mann Whitney Test (not the signed rank sum test for Paired data); ANOVA/Bonferroni; Kruskal Wallis test

Last time One way ANOVA F-test and Bonferroni shortcut comparison test

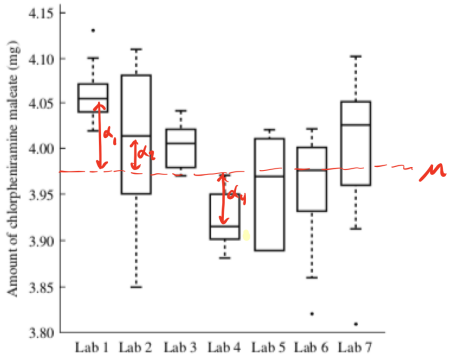


FIGURE 12.1 Boxplots of determinations of amounts of chlorpheniramir in tablets by seven laboratories.

J

I

model

overall mean; differential effect of i^{th} treatment

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0,\sigma^2)$$

$$SSB = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\bar{Y}_{i\cdot}-\bar{Y})^2 \quad \text{and} \quad \frac{SSB}{\sigma^2} \sim \chi^2_{I-1}$$

A nova

Assumptions of F-test

1) our data Y_{ij} is normal
2) the variance of each treatment group is the same σ^2
3) All observations are independent.

Today (1) Nonparametric Kruskal Wallis test — sec 12.2.3

(2) Matrix notation in multiple regression

(3) Simple standard regression model ($k=1$)

(4) Formula for regression coefficients

— sec 14.2

Sec 12.2.3 Kruskal Wallis (KW) test

$R_{ij} \in \{1, 2, \ldots, N\}$

R_{ij} = rank of Y_{ij} in the combined treatment grps.

Null: each treatment grp has same distribution
Alt: not all treatment grps have same dist.

let $SSB = \sum_{i=1}^{I} J_i (\bar{R}_{i\cdot} - \bar{R})^2$ be the between grp sum of squares

If the null is true that each treatment grp has the same distribution then SSB should be small.

As before $\frac{SSB}{\sigma_R^2} \sim \chi^2_{I-1}$

σ_R^2 ← can show that is $\frac{N(N+1)}{12}$

```
library(tidyr)
#time to failure for three types of engine bearings

treatment1 <- c(1,2,3,4,5)
treatment2 <- c(6,7,8,9)
treatment3 <- c(10,11,12)

#you should check that normality assumption is satisfied.
qqnorm(treatment1)
```

```
library(dplyr)
library(ggplot2)
```


b: nonparametric Kruskal-Wallis test

We can decide whether the treatments have an effect without assuming they are normal and have equal variances.

We test the hypothesis with a nonparametric method (Kruskal-Wallis test) which is similar to the Mann-Whitney rank sum test.

H_0: All the treatment distributions are the same.

H_1: At least one of the treatment distributions is different.

Step 1

We rank each observation without regard for treatment group, beginning with a rank 1 for the smallest observation. We compute the mean rank for each treatment and the mean of those means. Finally we square the difference of each treatment's average with the total mean.

```
df_tidy <- df_tidy %>% mutate(ranks=rank(treatments))
df_tidy
```

groups <fctr>	treatments <dbl>	ranks <dbl>
T1	1	1
T1	2	2
T1	3	3
T1	4	4
T1	5	5
T2	6	6
T2	7	7
T2	8	8
T2	9	9
T3	10	10

1-10 of 12 rows Previous 1 2 Next

```
mean_rank_df <- df_tidy %>% group_by(groups) %>% summarize(mean_rank=mean(ranks))
mean_rank_df
```

groups <fctr>	mean_rank <dbl>
T1	3.0
T2	7.5
T3	11.0

3 rows

```
joined_df <- df_tidy %>%
  left_join(mean_rank_df) %>%
  mutate(tot_mean=mean(mean_rank)) %>%
  mutate(diff_sq=(mean_rank-tot_mean)^2)
joined_df
```

groups <fctr>	treatments <dbl>	ranks <dbl>	mean_rank <dbl>	tot_mean <dbl>	diff_sq <dbl>
T1	1	1	3.0	6.5	12.25
T1	2	2	3.0	6.5	12.25
T1	3	3	3.0	6.5	12.25
T1	4	4	3.0	6.5	12.25
T1	5	5	3.0	6.5	12.25
T2	6	6	7.5	6.5	1.00
T2	7	7	7.5	6.5	1.00
T2	8	8	7.5	6.5	1.00
T2	9	9	7.5	6.5	1.00
T3	10	10	11.0	6.5	20.25

1-10 of 12 rows Previous 1 2 Next

Step 2

We compute the Kruskal-Wallis test statistic, `K`, to obtain a normalized measure ofhow much the average ranks within each treatment group deviate from the average rank of all the observations.

```
N <- sum(J)
N
```

```
## [1] 12
```

```
SSB <- sum(joined_df$diff_sq)  #sum of square between
SSB
```

```
## [1] 126
```

```
K <- (12/(N*(N+1)))*sum(joined_df$diff_sq)
K
```

```
## [1] 9.692308
```

```
1-pchisq(K,df=2)
```

```
## [1] 0.007858545
```

Using `Kruskal.test()`

We can also do a Kruskal-Wallis test with the built in kruskal.test() function in R.

```
treatments
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12
```

```
groups <- c(rep(1,5),rep(2,4),rep(3,3))
groups
```

```
##  [1] 1 1 1 1 1 2 2 2 2 3 3 3
```

```
kruskal.test(treatments~groups)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  treatments by groups
## Kruskal-Wallis chi-squared = 9.6923, df = 2, p-value = 0.007859
```

reject three

We accept the null that the two distributions are the same.

Previous Topic

(2)

Consider matrices:

$$y=\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad X_0=\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1} \quad X_1=\begin{bmatrix} x_{12} \\ \vdots \\ x_{n2} \end{bmatrix}_{n\times 1} \cdots X_k=\begin{bmatrix} x_{1k} \\ \vdots \\ x_{nk} \end{bmatrix}_{n\times 1} \quad \beta=\begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1} \quad e=\begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}_{n\times(k+1)} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} \beta_0+\beta_1 x_{11}+\cdots+\beta_k x_{1k}+e_1 \\ \vdots \\ \beta_0+\beta_1 x_{n1}+\cdots+\beta_k x_{nk}+e_n \end{bmatrix}_{n\times 1}$$

$X = [x_0, \ldots, x_k]_{n\times(k+1)}$ ← design matrix

$n \geq k+1$

X full rank

↳ $x_0, x_1, \ldots, x_k$ column vectors are independent

so $X'X$ is $(k+1)\times(k+1)$ invertible matrix

The standard statistical model is

$$y = X\beta + e$$

The e_i are independent RVs w/ $E(e_i)=0$, $\mathrm{Var}(e_i)=\sigma^2$

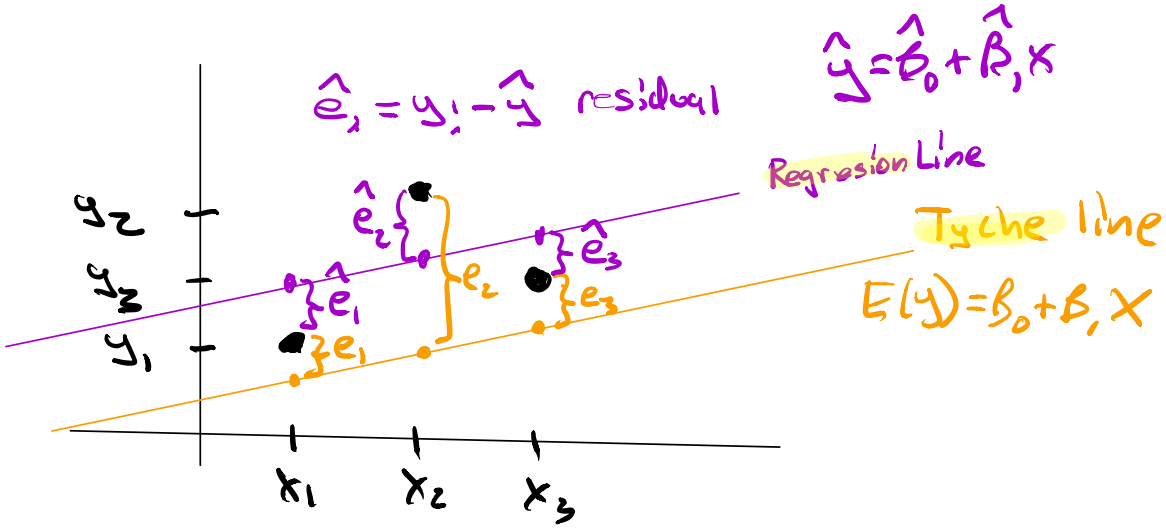
We say ϵ is <u>homoscedastic</u> since $\text{Var}(\epsilon_i)$ doesn't depend on X.

X: are fixed vectors

ex (simple standard statistical model)

$K=1$
$n=3$ Let's assume there is a linear relationship between x = father's height and y = son's height

$$y = \beta_0 + \beta_1 x + e$$

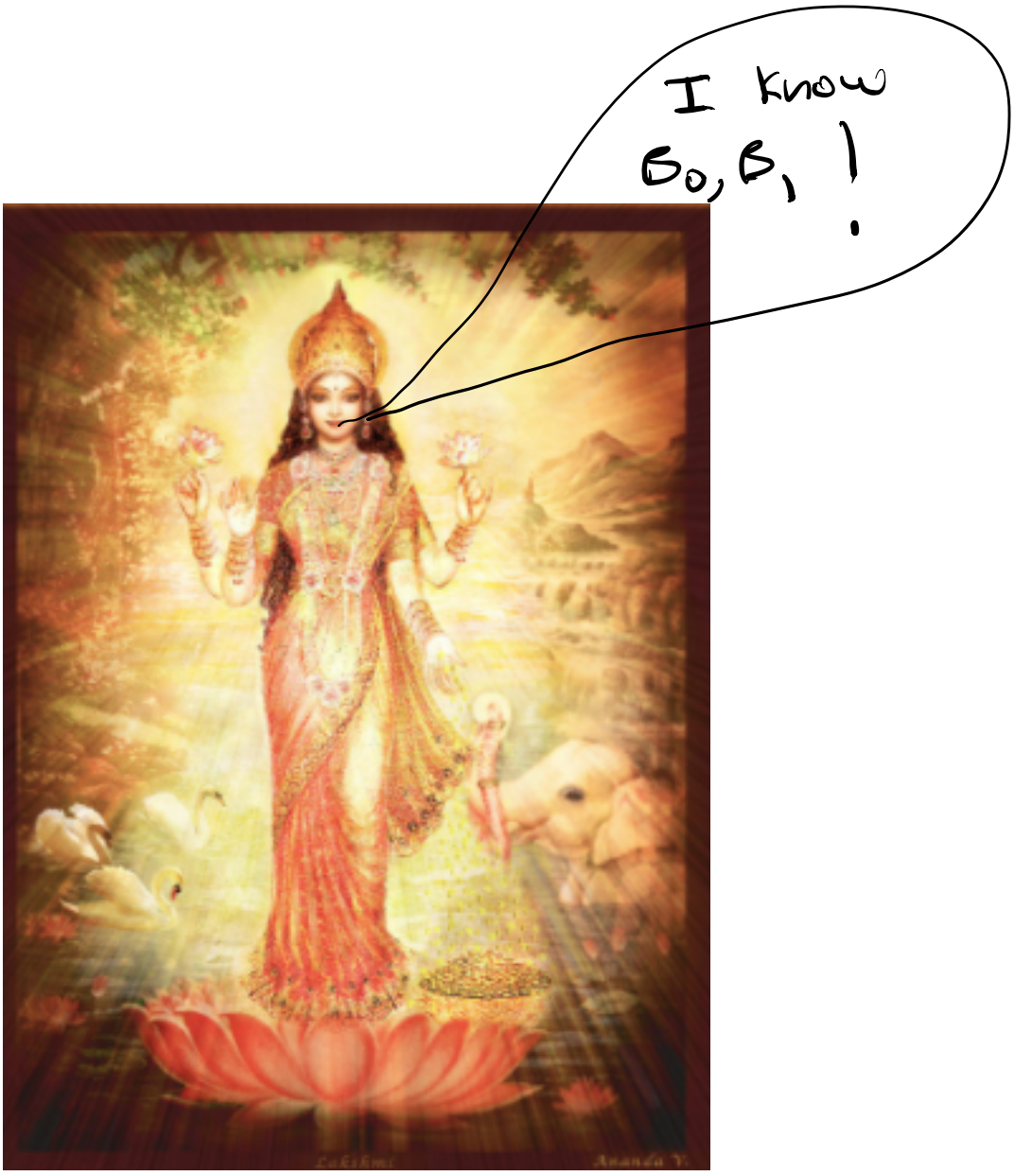


We wish to find the best fitting line through our data (regression line)

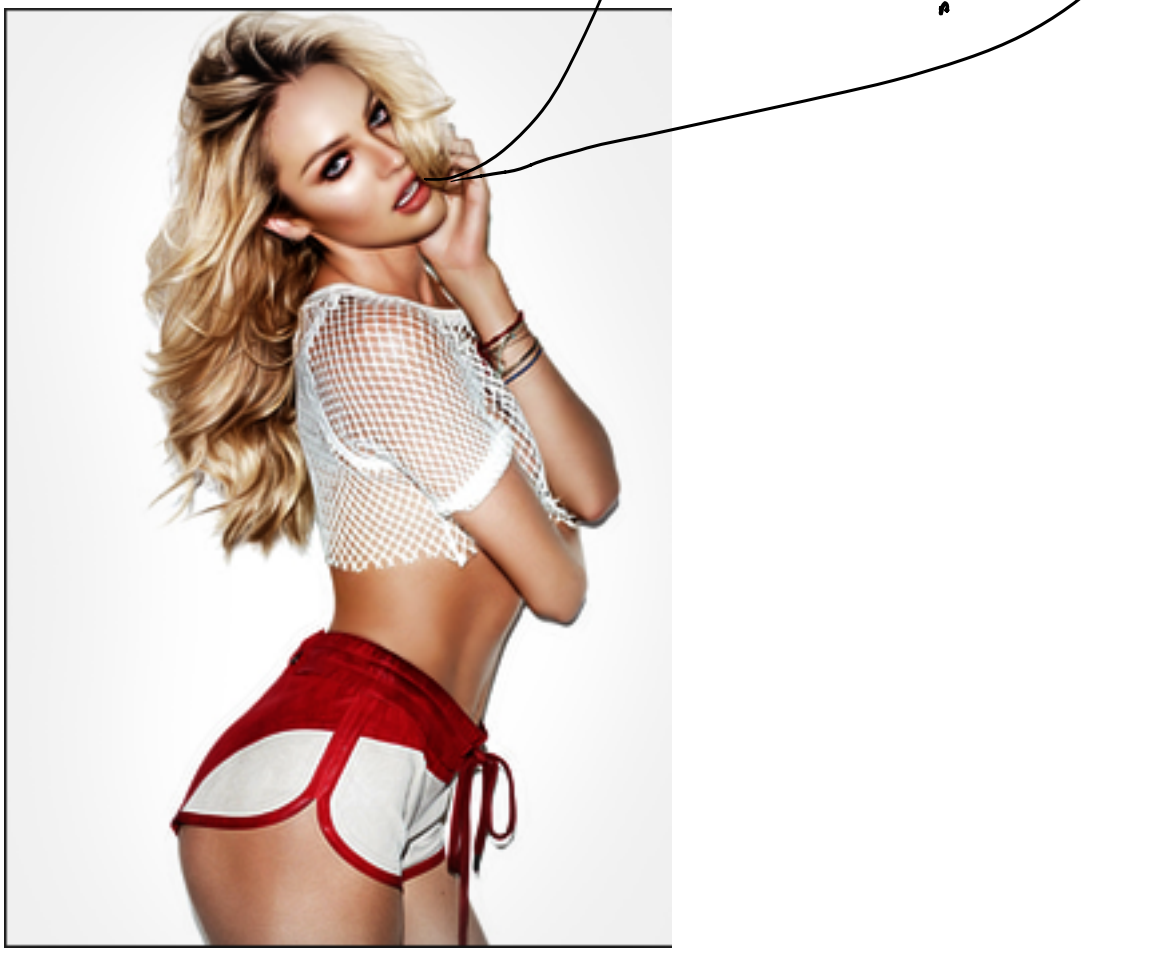
$\hat{e}_i = y - \hat{y}_i$ are the residuals

Since e is homoscedastic, the residuals $\hat{e}$ should also be approx homoscedastic

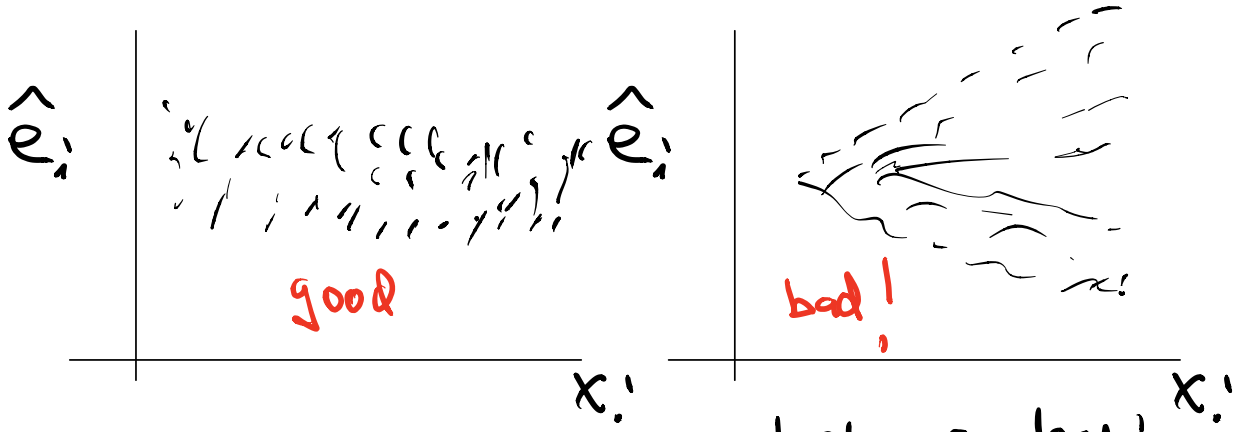
Image of Tyche the goddess of fortune.
Only Tyche knows B_0, B_1.



An image of a modern day Tyche from Sin City.



You should always graph $\hat{e}_i$, called a residual plot. The residual spread should be constant



heteroscedastic residual plot.

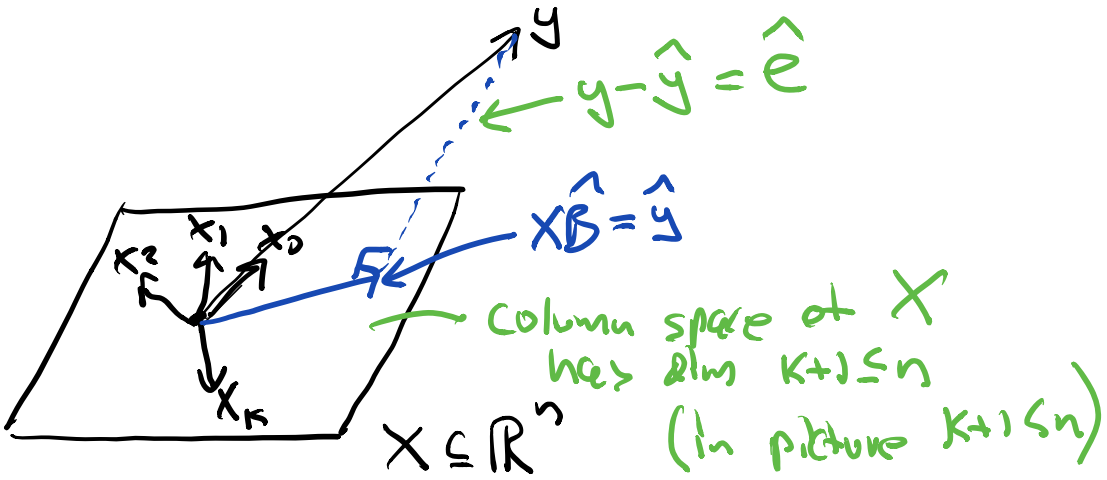
We want to minimize the sqrt of sum of squares of the residuals $\sqrt{\sum_{i=1}^{n} \hat{e}_i^2}$

$\|\hat{e}\|$ Euclidean norm of $\hat{e} = \begin{bmatrix} \hat{e}_1 \\ \vdots \\ \hat{e}_n \end{bmatrix}$

(3) Geometric derivation of the regression coeffs.

$y \in \mathbb{R}^n \quad (n \geq K+1)$

The columns of X span a $K+1$ dimensional subspace since X is full rank.



Find $\hat{B}$ such that $\|y - X\hat{B}\| = \|y - \hat{y}\| = \|\hat{e}\|$ is as small as possible.

Take the orthogonal projection of y on the column space of X.

$$\underset{(K+1)\times n}{X'} \cdot \underset{n\times 1}{\left(y - X\hat{\beta}\right)} = \underset{(K+1)\times 1}{0}$$

$$\begin{matrix} x_0 \\ x_1 \\ \vdots \\ x_K \end{matrix}\begin{pmatrix} \\ \\ \\ \\ \\ \end{pmatrix} \cdot \Big| = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(K+1)\times 1}$$

$$\Rightarrow X'y - X'X\hat{\beta} = 0$$

$$\Rightarrow X'y = X'X\hat{\beta}$$

$$\boxed{\underset{(K+1)\times 1}{\hat{\beta}} = (X'X)^{-1}X'y}$$

if y lies in column space of X

$$y = X\hat{\beta} \Rightarrow X'y = X'X\hat{\beta}$$

$$\Rightarrow \hat{\beta} = (X'X)^{-1}X'y \quad \text{as before.}$$

Next time for $K=1$ case

will show

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} E(y) - \hat{\beta}_1 E(x) \\ \dfrac{Cov(x,y)}{Var(x)} \end{bmatrix}$$

Last time

Stat 135 lec 32

Chap 14 Linear regression

indep var	cat	cat	quant	quant
dep var	quant	cat	quant	cat
	t-test (≤ 2 grps) ANOVA (≥ 2 grps)	χ^2 test	linear regression	logistic regression (didn't cover)
example	mean wt. of diff. treatment grps.	no. patients of diff. treatment grps.	predict ht. from wt.	predict binary outcome from wt.
chapters	11,12	9.5,13	14	

consider matrices:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad X_0 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1} \quad X_1 = \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix}_{n\times 1} \quad \cdots \quad X_k = \begin{bmatrix} x_{1k} \\ \vdots \\ x_{nk} \end{bmatrix}_{n\times 1} \quad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1} \quad e = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}$$

$$X = \underbrace{\begin{bmatrix} X_0, \cdots, X_k \end{bmatrix}}_{n\times(k+1)}$$

design matrix

where $n \geq k+1$ and X full rank

std statistical model: $y = X\beta + e$

We require e to be homoscedastic (i.e independent of x).

Properties of linear regression related to the variance of $\hat{\beta}$ relies on this.

Only Tycho knows β and e.

We see the data and can make a regression "line" $\hat{y} = \hat{\beta}X$ that minimizes the Euclidean norm of the residual vector $\hat{e} = \hat{y} - y$

We found $\hat{\beta} = (X'X)^{-1}X'y$

Today

1. Sec 14.3 simple regression

$$\hat{\beta}_1 = \frac{\text{Cov}(x,y)}{\text{var}(x)}$$

$$\hat{\beta}_0 = E(y) - \hat{\beta}_1 E(x)$$

2. in-class R exercise
3. sec 14.4 Statistical properties of $\hat{\beta}$.

① Sec 14.3

$$\hat{B} = (X'X)^{-1}X'Y$$

In the case the case of simple linear regression we will show

$$\hat{B}_1 = \frac{Cov(x,y)}{Var(x)}$$

$$\hat{B}_0 = \bar{E}(y) - \hat{B}_1 \cdot E(x)$$

$X = Unif(x_1, \ldots, x_n)$ ← x_n ← x_{n-1}

$Y = Unif(y_1, \ldots, y_n)$

$E(x) = \bar{x}$

$E(y) = \bar{y}$

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

EXAMPLE B Returning to Example A on fitting a straight line, we have

$\hat{B}_{2\times 1} = (X'X)^{-1}X'y$ where

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

Confirm that

$\hat{B}_1 = \dfrac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

Thus,

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$$= \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

$$= \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{bmatrix} \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right) \\ n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \end{bmatrix}$$

$n^2\mathrm{Var}(x)$ → denominator; $n^2\mathrm{Cov}(x,y)$ ← second entry

which agrees with the earlier calculation. ■

so $\boxed{\hat{B}_1 = \frac{\text{cov}(x,y)}{\text{var}(x)}}$ ✓

We show that $\hat{B}_0 = E(y) - \hat{B}_1 E(x)$.

$$\Sigma y_i \Sigma x_i^2 - \Sigma x_i \Sigma x_i y_i = \Sigma y_i \Sigma x_i^2 - \Sigma y_i (\Sigma x_i)^2 - \Sigma x_i \Sigma x_i y_i + \Sigma y_i (\Sigma x_i)^2$$

$$= \Sigma y_i \left[\Sigma x_i^2 - (\Sigma x_i)^2 \right] - \left[\Sigma x_i y_i - \Sigma x_i \Sigma y_i \right] \Sigma x_i$$

$$= n^2 E(y)\,\text{var}(x) - n^2 \text{cov}(x,y) E(x)$$

Hence

$$\hat{B}_0 = \frac{\Sigma y_i \Sigma x_i^2 - \Sigma x_i \Sigma x_i y_i}{n \Sigma x_i^2 - (\Sigma x_i)^2} = \frac{n^2 E(y)\text{var}(x) - n^2 \text{cov}(x,y) E(x)}{n^2 \text{var}(x)}$$

$$= E(y) - \underbrace{\frac{\text{cov}(x,y)}{\text{var}(x)}}_{=\hat{B}_1} E(x)$$

so $\boxed{\hat{B}_0 = E(y) - \hat{B}_1 E(x)}$ ✓

Next

Find $\hat{B}$ using Calculus approach.

$S(\beta_0, \beta_1) = \sum (y_i - \beta_0 - \beta_1 x_i)^2$

want to find β_0, β_1 that minimize this.

To find β_0 and β_1, we calculate

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

$\hat{e}_i$

$\sum \hat{e}_i = 0$

$\sum x_i \hat{e}_i = 0$

called normal equations

Setting these partial derivatives equal to zero, we have that the minimizers $\hat{\beta}_0$ and $\hat{\beta}_1$ satisfy

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

Solving for $\hat{\beta}_0$ and $\hat{\beta}_1$, we obtain

$$\hat{\beta}_0 = \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Problem 10 at the end of the chapter asks you to derive the following useful equivalent expressions:

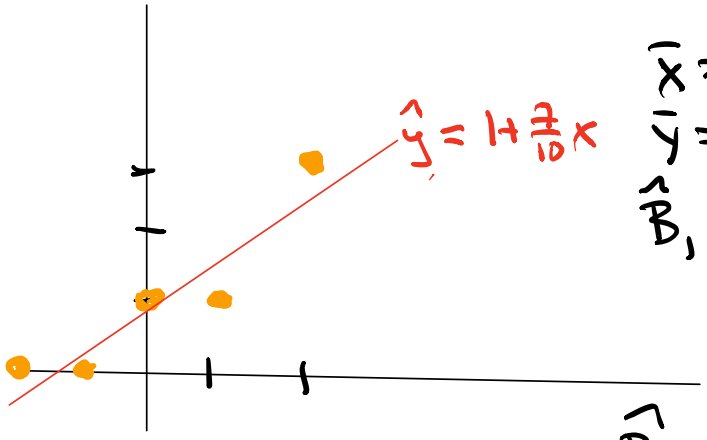
$\bar{y} - \hat{\beta}_1 \bar{x}$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$\frac{\text{Cov}(x,y)}{\text{Var}(x)}$

Ex Use the method of least squares to fit $n=5$ data pts $(-2,0), (-1,0), (0,1), (1,1), (2,3)$



$\bar{x} = 0$

$\bar{y} = 1$

$\hat{\beta}_1 = \frac{\text{cov}(x,y)}{\text{var}(x)} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

$= \frac{7}{10}$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 1 - \frac{7}{10} \cdot 0 = 1$

$\boxed{\hat{y} = 1 + \frac{7}{10}x}$

On HW 9 you will show the regression line can be written

$$\frac{\hat{y} - \bar{y}}{\sigma_y} = r \left(\frac{x - \bar{x}}{\sigma_x} \right)$$

r: Correlation coefficient of x, y.

Stat135 lecture 18

Simple Linear Regression

Start Over

in-class exercise lec 33 on bCourses.

Simple Linear Regression

Here is a basic example of how to draw a linear regression line in R

same data as in above example.

```
x <- -2:2
y <- c(0,0,1,1,3)
df <- data.frame(x,y)
df
```

x <int>	y <dbl>
-2	0
-1	0
0	1
1	1
2	3

5 rows

y ~ x

```
reg <- lm(formula=y~x)
summary(reg)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##          1          2          3          4          5
##  4.000e-01 -3.000e-01 -2.776e-16 -7.000e-01  6.000e-01
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0000     0.2708   3.693   0.0345 *
## x             0.7000     0.1915   3.656   0.0354 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6055 on 3 degrees of freedom
## Multiple R-squared:  0.8167, Adjusted R-squared:  0.7556
## F-statistic: 13.36 on 1 and 3 DF,  p-value: 0.03535
```

```
reg$coefficients
```

```
## (Intercept)           x
##         1.0         0.7
```

Lets find the expected y value for an x value of 3:

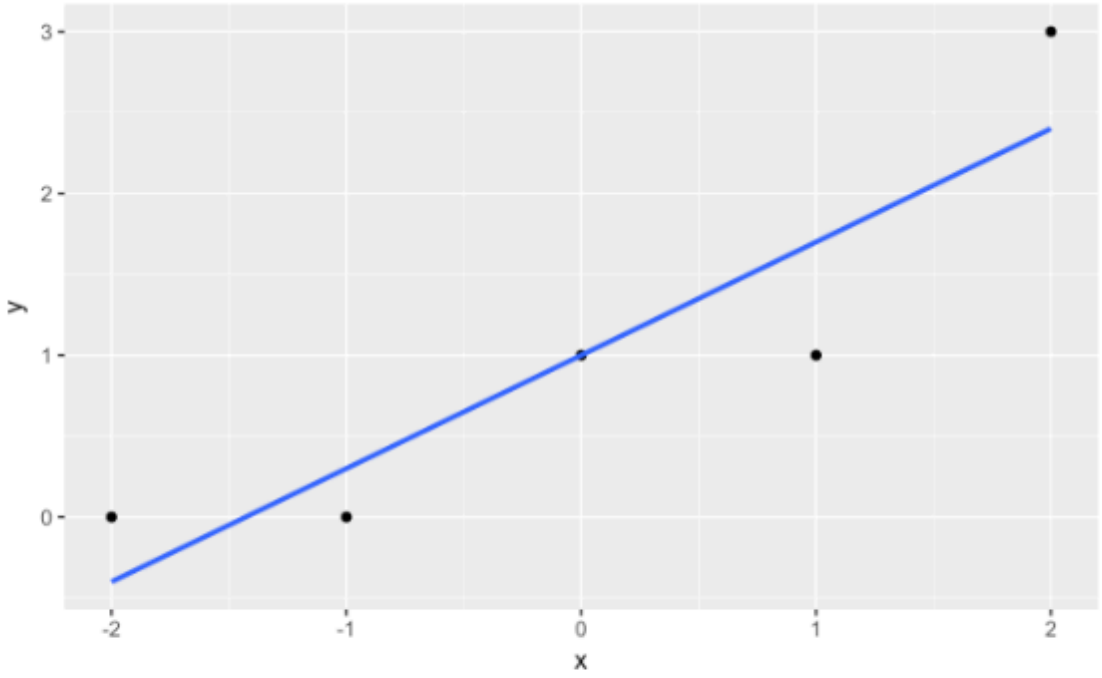
"1" "7"

```
reg$coefficients[1]+ 3*reg$coefficients[2]
```

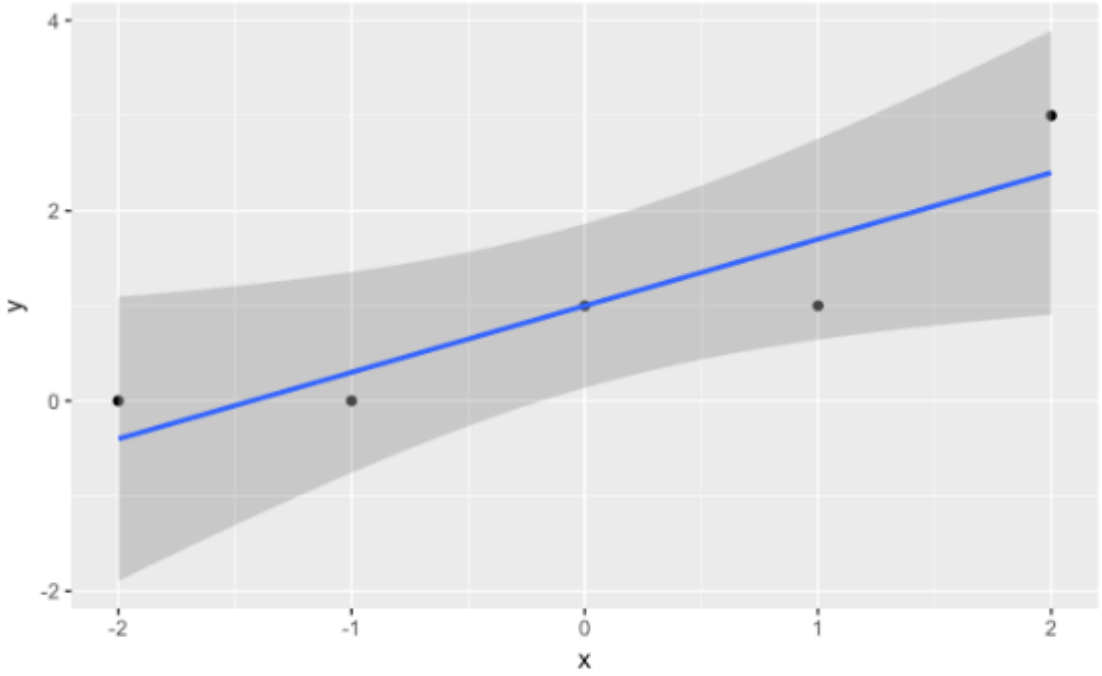
= 1+3(.7)

```
## (Intercept)
##         3.1
```

```
df %>% ggplot(aes(x=x,y=y)) + geom_point() + geom_smooth(method=lm,se=FALSE)
```




```
df %>% ggplot(aes(x=x,y=y)) + geom_point() + geom_smooth(method=lm)
```



example of drawing a regression line

There are two variables age and height. ages: 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29
heights: 76.1, 77, 78.1, 78.2, 78.8, 79.7, 79.9, 81.1, 81.2, 81.8, 82.8, 83.5

Find the equation of the regression line predicting height for different ages (i.e. x=ages, y=heights).
Predict the height of a 1 year old. Draw the regression line and the points (no error bars).

Code | Start Over | Solution | Run Code

```
1
2
3
```

Sec 14.4 Statistical properties of least square estimators

A vector valued RV is called a random vector.

ex $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

We can generalize covariance of a random variable to covariance of a random vector.

let $\hat{B} = \begin{bmatrix} \hat{B}_0 \\ \vdots \\ \hat{B}_k \end{bmatrix}$ be random vector

$$E(\hat{B}) = \begin{bmatrix} E(\hat{B}_0) \\ \vdots \\ E(\hat{B}_k) \end{bmatrix}$$

$$\text{Var}(\hat{B}) = \begin{bmatrix} & \text{Cov}(\hat{B}_i, \hat{B}_j) & \end{bmatrix} \quad \begin{matrix} i = 0, \dots, k \\ j = 0, \dots k \end{matrix}$$

$(k+1) \times (k+1)$

$\text{Var}(\hat{B})$ in Rice $= \Sigma_{\hat{B}\hat{B}}$

Fact Thm A, B p 573, 574

$$E(\hat{B}) = B$$

$$\text{Var}(\hat{B}) = \sigma^2 (X'X)^{-1}$$

P574 Rice

EXAMPLE A We return to the case of fitting a straight line. From the computation of $(\mathbf{X}^T\mathbf{X})^{-1}$ in Example B in Section 14.3, we have

$$\Sigma_{\hat{\beta}\hat{\beta}} = \frac{\sigma^2}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}$$

Therefore,

algebra see below*

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{n\,\mathrm{var}(x)}\right]$$

$n^2\,\mathrm{var}(x)$

$$\mathrm{Var}(\hat{\beta}_1) = \frac{n\sigma^2}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \frac{\sigma^2}{n\,\mathrm{var}(x)}$$

$n^2\,\mathrm{var}(x)$

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \frac{-\sigma^2 \bar{x}}{n\,\mathrm{var}(x)}$$

■

* algebra:

$$\text{var}(x) = \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

$$\Rightarrow \sum x_i^2 = n\,\text{var}(x) + n\bar{x}^2$$

$$\Rightarrow \text{var}(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{n^2\,\text{var}(x)} = \sigma^2\left(\frac{n\,\text{var}(x) + n\bar{x}^2}{n^2\,\text{var}(x)}\right)$$

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{n\,\text{var}(x)}\right)$$

Knowing the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ will allow us to do hypothesis testing on whether x and y are related and make confidence intervals.

We don't know σ^2 so next time we will find an estimator of σ^2.