Chapter 20

Comparing Means between two Paired Groups
DEPENDENT Groups

Twin Pairs	IQ OLDER	IQ Younger	Difference younger − older
1	115	120	5
2	87	90	3
3	104	101	−3
4			
⋮			
20	136	132	−4

Sample mean	$\bar{y}_1$	$\bar{y}_2$	$\bar{d} = \bar{y}_1 - \bar{y}_2$
Sample SD	s_1	s_2	s_d

n pairs

$$\bar{d} = \frac{d_1 + d_2 + \cdots + d_n}{n} \qquad \bar{d} = \bar{y}_1 - \bar{y}_2$$

$$s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}}$$

Inference for μ	Inference for mean pop. difference d
mean μ ; $SD = \sigma$	mean $= \mu_{diff.}$ $SD = \sigma_{diff}$
statistic $\bar{y}$	statistic $\bar{d} = \bar{y}_1 - \bar{y}_2$
sampling dist. of $\bar{y}$: $N(\mu, \frac{\sigma}{\sqrt{n}})$	sampling dist. of $\bar{d}$: $N(\mu_d, \frac{\sigma_d}{\sqrt{n}})$
when σ is unknown, we estimate σ with s use the t model	when σ_d is unknown we estimate σ_d with s_d and use the t model

Confidence Interval for the true mean difference μ_d is:

$$\bar{d} \pm t^*_{n-1} \frac{s_d}{\sqrt{n}}$$

where the critical value t_{n-1}^{*} we get from the t table

$\Rightarrow$ This Confidence Interval can be viewed as a one sample t-interval for the Population mean difference μ_d.

t model Assumptions

1) random sample of n pairs

2) The pairs are independent of each other, but the subjects within each pair are of course DEPENDENT

# of pairs $<$ 10% of the population size.

3) d_i's (differences) distributed Normally with no restriction on sample size n

Symmetric + Unimodal

If the d_i's have unknown distribution or Non-Normal dist. then for large

enough sample size $n \geq 30$ we can apply the CLT as well as the t model.

1) Pop. SD of d_i's, $\sigma_{diff.}$ is unknown $\Rightarrow$ USE t distribution
2) The data must be paired

The paired t-test is like a one sample t-test on the population mean difference μ_d

H_0: $\mu_d = \Delta_0$

H_a: $\mu_d \neq \Delta_0$

$\mu_d > \Delta_0$

$\mu_d < \Delta_0$

Null model & test statistic

One sample t-test for μ

$H_0: \mu = \mu_0$ σ unknown

$$\bar{y} : N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right)$$

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

t model: $df = n-1$

Paired t-test for a population mean difference μ_d

σ unknown

$H_0: \mu_{diff.} = \Delta_0$

$$\bar{d} : N\left(\Delta_0, \frac{\sigma_d}{\sqrt{n}}\right)$$

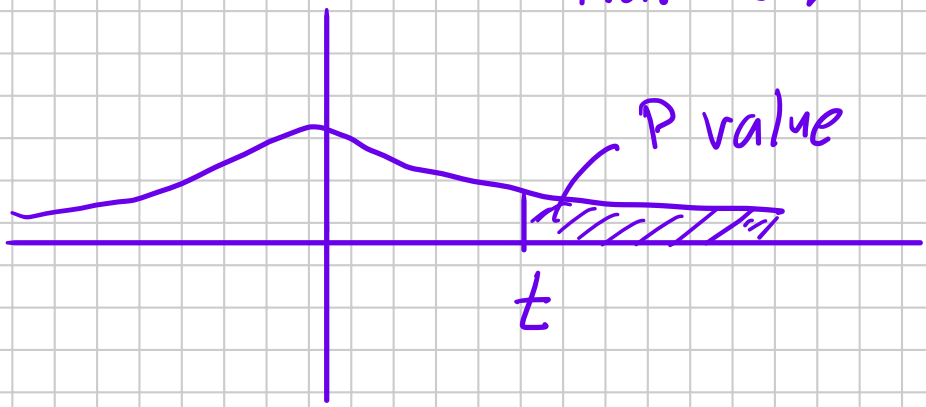
$$t = \frac{\bar{d} - \Delta_0}{s_d/\sqrt{n}}$$

Use t model $df = n-1$

P value

$H_0: \mu_d = 0$

$H_a: \mu_d > 0$



$P < \alpha$ reject null

$P \geq \alpha$ don't reject null

Exercise Paired t test

Patient	Existing Drug	New drug	diff = New drug − old drug
1	3.5	2.6	diff = 2.6 − 3.5 = −0.9
2	2.6	2.8	d = 0.2
3	3.0	3.1	d = 0.1
4	1.9	2.4	d = 0.5
5	2.9	2.9	d = 0
6	2.4	2.2	d = −0.2
7	2.0	2.1	d = 0.1

90% C.I for μ_d

$$\bar{d} \pm t_{(n-1)} \frac{s_d}{\sqrt{n}}$$

$$-0.029 \pm 1.943 \frac{(0.439)}{\sqrt{7}}$$

$\bar{d} = -0.029$

$s_d = 0.439$

$df = 7 - 1 = 6$

$t^*_{df=6} = 1.943$

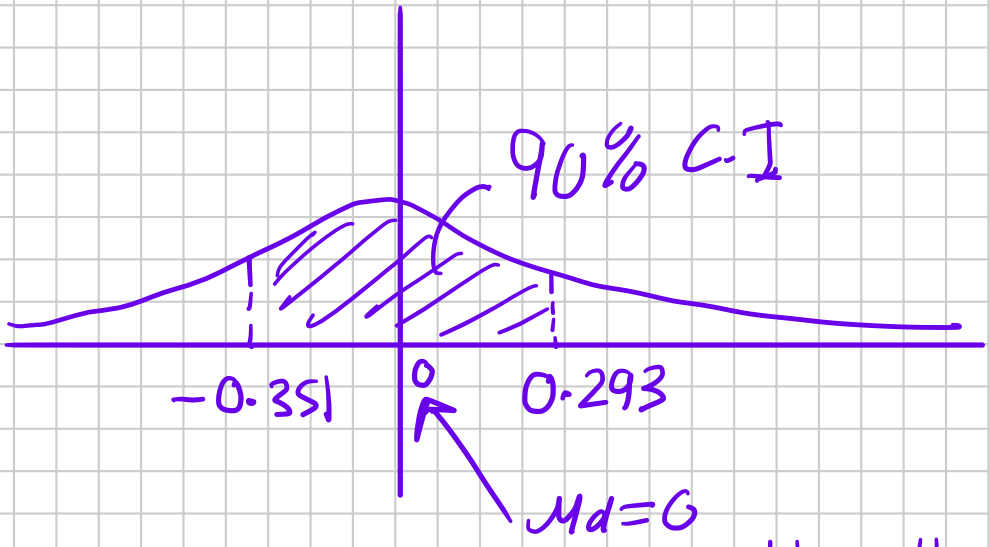
-0.029 ± 0.322

$-0.029 - 0.322 > -0.029 + 0.322$

$(-0.351 > 0.293)$ 90% C.I

b) $H_0: \mu_{diff} = \mu_{new} - \mu_{old} = \boxed{0}$

$H_a: \mu_{diff} \neq 0$ $\alpha = 0.10$



90% C.I agrees with Hypothesis test with $\alpha = 0.10$, the value of $\mu_d = 0$ is inside the 90% C.I

∴ we cannot reject the null hypothesis with $\underline{\alpha = 0.10}$

b) Hypothesis test

$H_0: \mu_d = \mu_{new} - \mu_{old} = 0$

$H_a: \mu_d \neq 0$

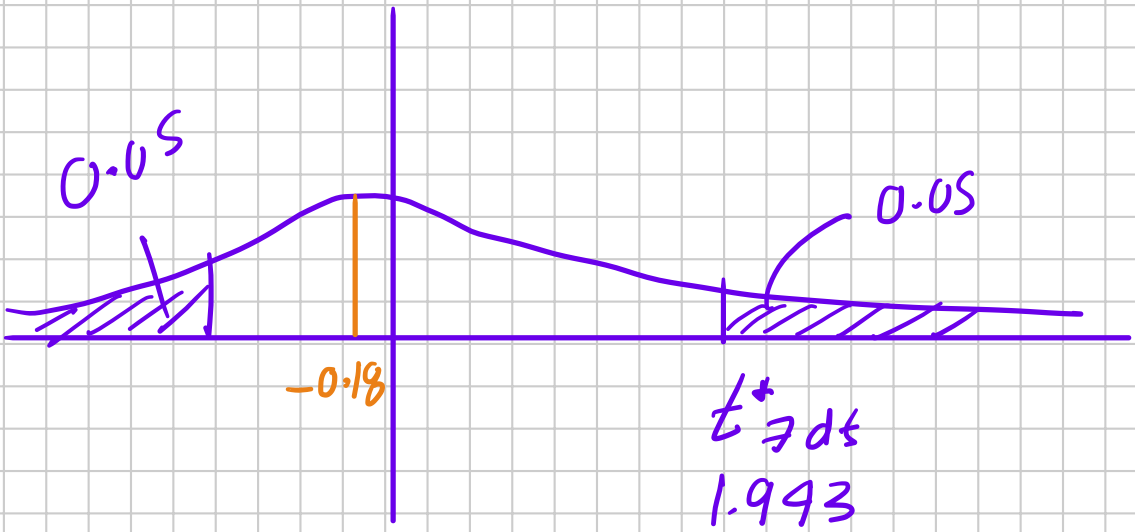
$\bar{d} = -0.029$

$s_d = 0.439$

$n = 7$

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} =$$

$$t = \frac{-0.029 - 0}{0.439/\sqrt{7}} = -0.175$$



$t = -0.18$ doesnot fall in rejection region. ∴ Cannot reject null hypothesis.

Ch. 22 Comparing Counts

Contingency table we examined the relationship between two categorical variables and we use conditional probabilites to explore the association between these two categorical variables but now we have the tools to explore this association more formally by doing a hypothesis test

	Having Heart Disease YES	NO	Total
High Cholest	11	4	15
Low Choles	2	6	8
Total	13	10	23

Is there an assoc. between Diet type and Heart Disease?

Conditional Distribution for having Heart Disease for High cholestrol diet

Heart Disease for all with High Cholestrol		TOTAL
YES	NO	
11	4	15
11/15 = 73.3%	4/15 = 26.7%	

$$\frac{73.3\%}{26.7\%} = \boxed{2.74}$$

If a person has a High Cholestrol diet then they are 2.74 times more likely to have heart disease

---

Conditional distribution of heart disease for a low cholestrol diet

Heart Disease given a low chol. diet

Having Heart Disease YES	NO	Total
2	6	8
$\frac{2}{8} = 25\%$	$\frac{6}{8} = 75\%$	100%

73.3% of people with High Chol. diet had heart disease whereas only 25% of those with low Chol. diet had heart disease

This suggests diet type and having heart disease are associated

∴ Dependent

How Big a difference between the Conditional distributions is required to declare an association

1) No difference ⇒ Independence
No Association

2) Very Different ⇒ dependence/
association

3) between (1) & (2) ⇒ need formal
statistical

EXAMPLE

	Drinks per month			
	Abstain	1-60	over 60	Total
Single	67	213	74	354
Married	411	633	129	1173
Widowed	85	51	7	143
Divorced	27	60	15	102
Total	590	957	225	1772

Question: Are Marital Status and alcohol consumption
Independent or Is there an association

Conditional dist. of alcohol consumption for Single individuals

for Given

SINGLE

Abstain	1–60	over 60	Total
67	213	74	354
(18.93%)	(60.17%)	(20.9%)	(100%)

Conditional dist. of alcohol consumption for (Given) married individuals.

Abstain	1–60	over 60	Total
411	633	129	1173
$\frac{411}{1173} = 35.04\%$	53.96%	11%	100%

$$\frac{20.9\%}{11\%} = 1.9 \approx 2$$

Single people are twice as likely to drink heavily $\Rightarrow$ over 60 drinks per month

Intuitively we can see that alcohol consumption is associated with marital status (Compare the percentages)

18.93% of single individuals abstain

35.04% of married ~~/~~ abstain

Distribution of alcohol consumption is not the same for different marital status groups.

Conditional distribution of marital status for (given) those who drink more than 60 drinks per month

Column %

Over 60 drink	Single	Married	Widowed	Divorced	Total
	74	129	7	15	225
	74/225	129/225	7/225	15/225	100%
	32.9%	57.33%	3.11%	6.67%	

57.33% of all the heavy drinkers are married people whereas 32.9% of all the heavy drinkers are single people

This is not useful since there the married people sample size is almost 3 times as large as single people.

Formal Testing of Independence of 2 categories

H_0: The 2 categories are Independent or not associated

Diet and heart disease are not associated

H_a: the 2 categories are Associated / Related / Dependent

Row variable has r categories

eg. marital status $r=4$

Column Variable has C categories

eg. Drinking Frequency $C=3$

The total number of counts is n

eg. 1772 people

In constructing the test statistic always assume that the null hypothesis

is correct.

Under the null hypothesis, no assoc. between rows and columns
Cond. dist. of the row variable will be identical across the categories of the column variable.

Handedness

	Right-handed	Left handed	Total
Male	54 = 60% of 90	6 = 60% of 10	(60)
Female	36 = 40% of 90	4 = 40% of 10	40
Total	(90)	10	(100)

Males are 60% of all people
Females " 40% " "

So we expect 40% of all right handed people to be female

To find expected values assuming no association between the categories

$$\frac{\text{row sum} \times \text{Column Sum}}{\text{total}}$$

Hypothesis Test Chi-Square

H_0: No association between the 2 Categories ie) rows and columns

$$E_{ij} = \frac{\text{column total}(j) \times \text{row total}(i)}{n}$$

$i = 1, 2, \ldots r$ (# of rows) $\quad j = 1, 2, \ldots C$ (# of columns)

Don't round expected values to integers when computing test statistic

Chi-Square Test

$$\chi^2 = \sum \frac{(\text{OBSERVED} - \text{Expected})^2}{\text{Expected}}$$

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$ sum is

taken across all r x c cells in the contingency table

O_{ij} is observed count in (i, j) cell.

O_{12} =

↓ ↘

row 1 Column 2

For All Hypothesis Test, the Big Picture is:

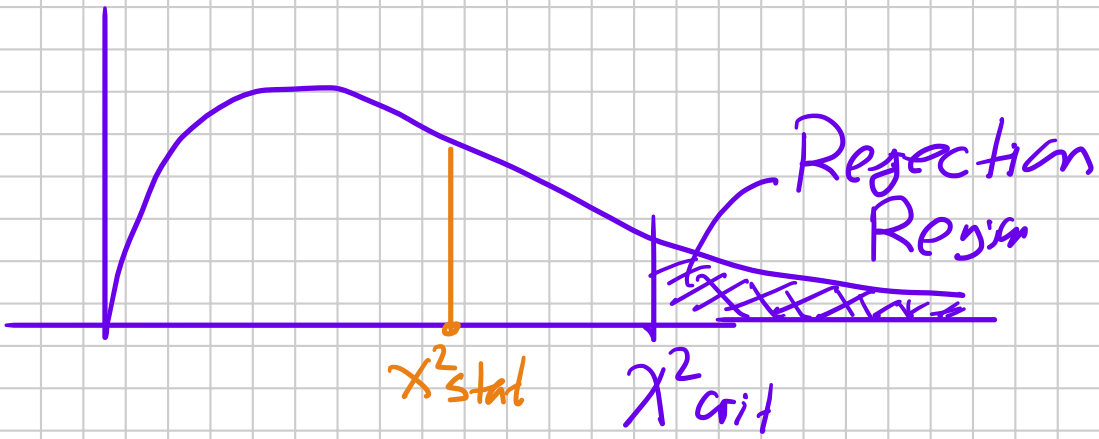
Large Z $\Rightarrow$ Reject Null

large t $\Rightarrow$ Reject $=$

large χ^2 $=$ $=$ $=$

A large χ^2 indicates disagreement between the expected and observed counts. This presents evidence against the null hypothesis

The χ^2 test is <u>always</u> right tailed



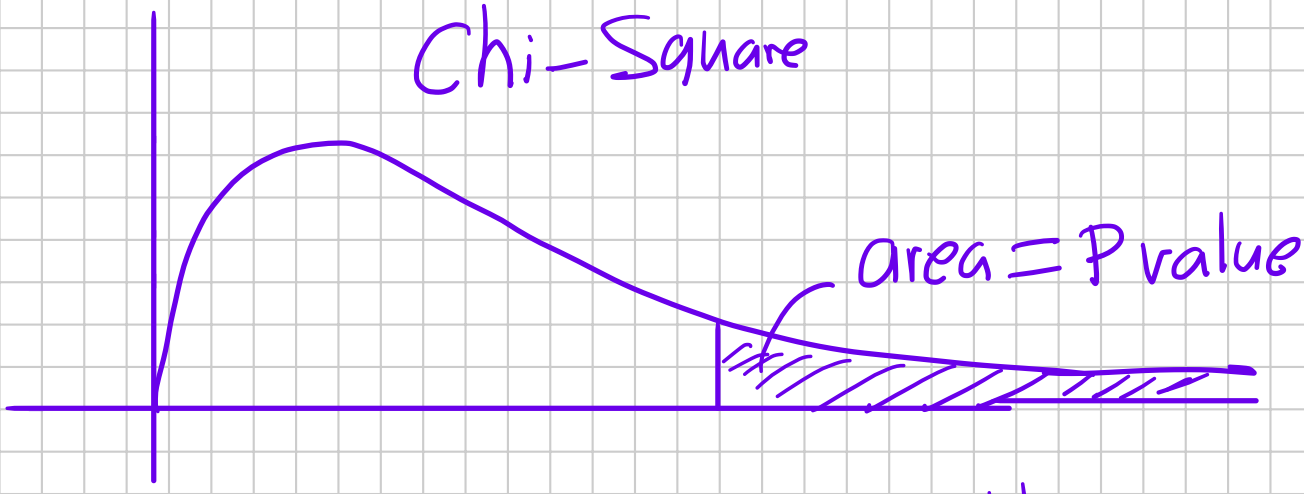
Under the null $\chi^2 \sim \chi^2_{df}$

$$df = (\text{rows} - 1)(\text{columns} - 1)$$

$$df = (r - 1)(c - 1)$$

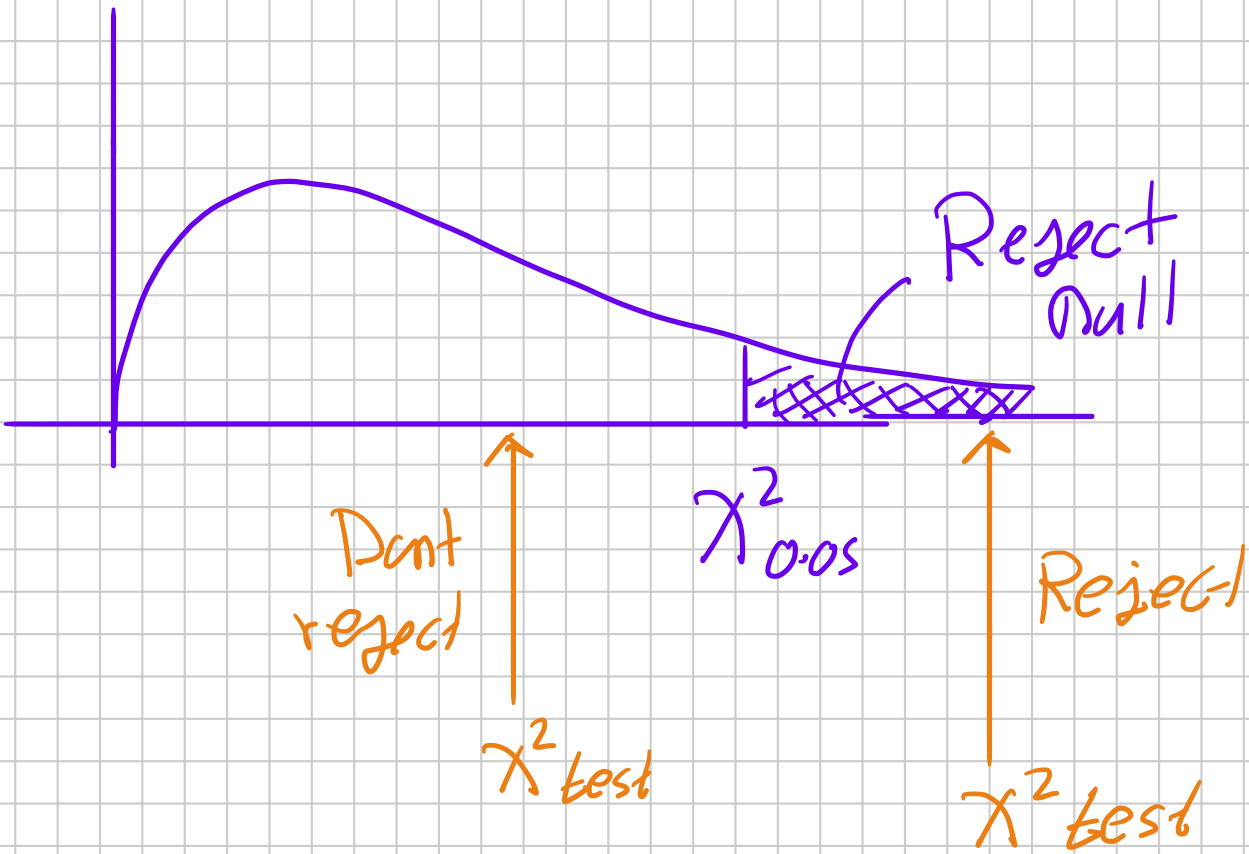
<u>Condition</u> each expected cell count has to be at least 5

χ^2 test stat



$P < \alpha$ reject null

$\Rightarrow$ There is an association between rows and columns



Reject H_0 $P < \alpha$

or $\chi^2_{stat} > \chi^2_{(r-1)(c-1)\,5\,\alpha}$

Don't reject H_0 $P \geq \alpha$

$\chi^2_{stat} \leq \chi^2_{(r-1)(c-1)}$

When H_0 is Rejected then we conclude that the two

Categorical variables are associated

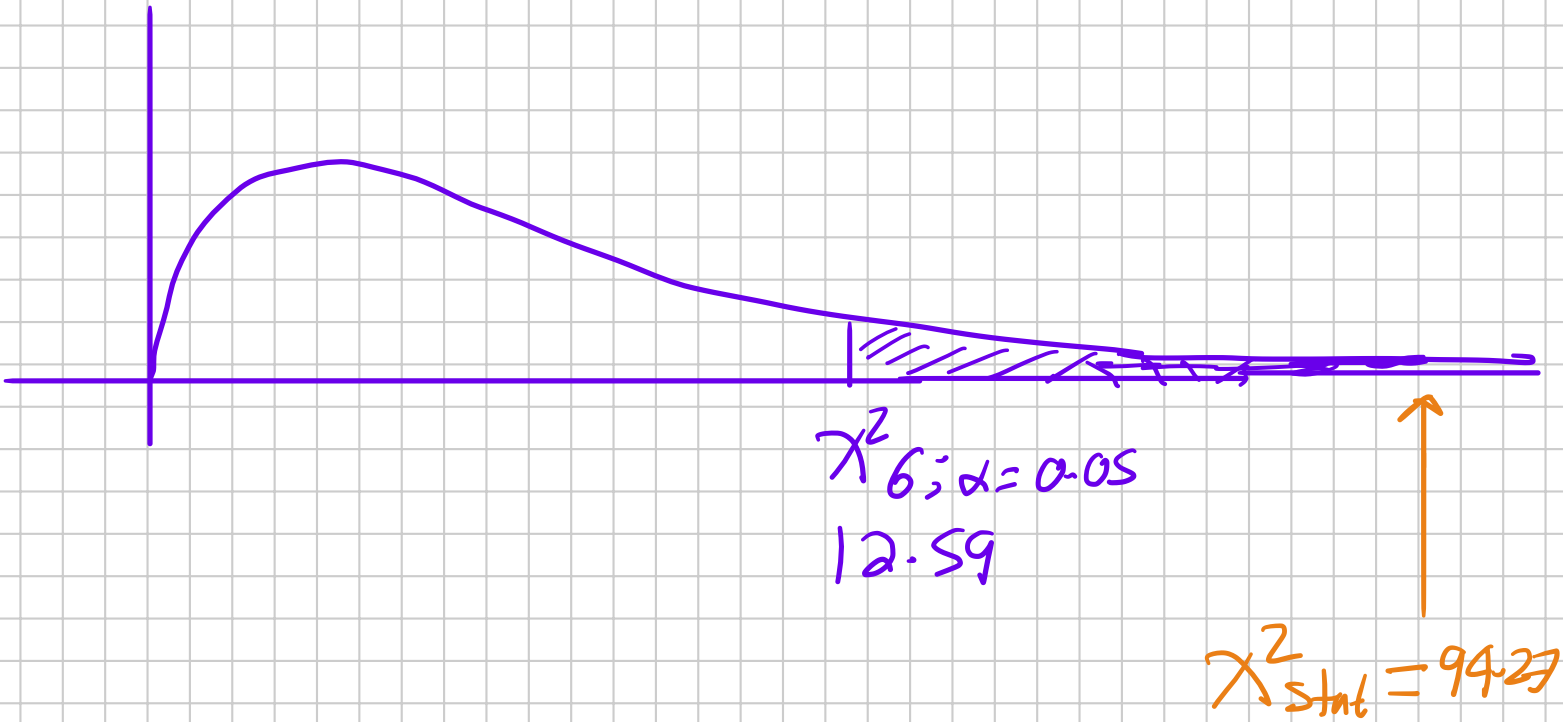
Marital Status + Drinking

$$\chi^2_{stat} = \sum \frac{(obs - Exp)^2}{Exp} = 94.269$$

$$df = (r-1)(c-1) = (4-1)(3-1) = \boxed{6}$$

at least 5 in each cell ✓

$$\chi^2_{6(\alpha=0.05)} = 12.59$$



$\therefore$ Since χ^2 stat $= 94.27 > \chi^2_{crit} = 12.59$

We reject null hypothesis and conclude that there is an association between Marital Status and alcohol consumption.