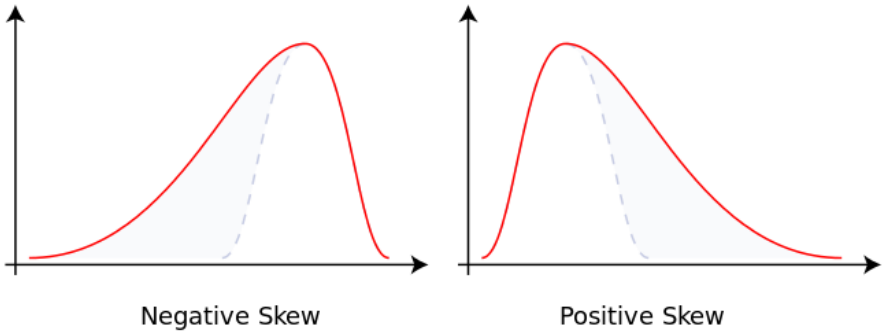
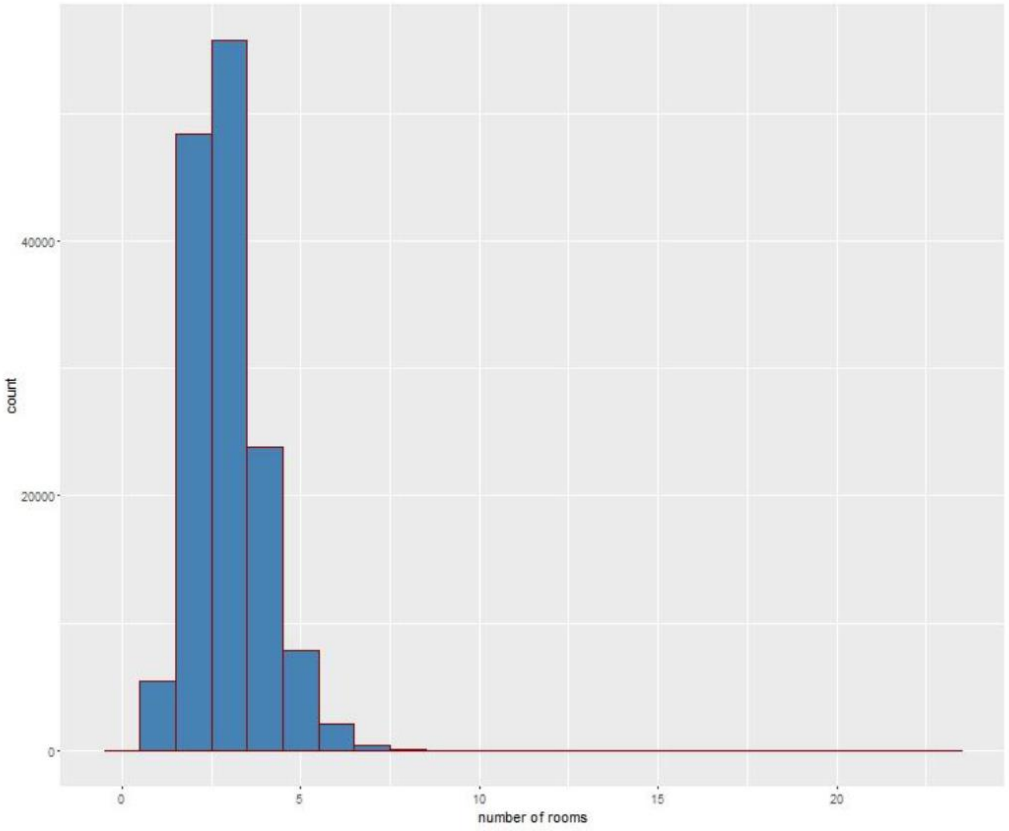
Q1

- یک متغیر عددی را انتخاب کرده و هیستوگرام مربوط به آن را ترسیم کرده، madality و Skewnessآن را تحلیل کنید. برای همین متغیر boxplot را ترسیم کرده و تعداد outlier ها را مشخص کنید.

Q1



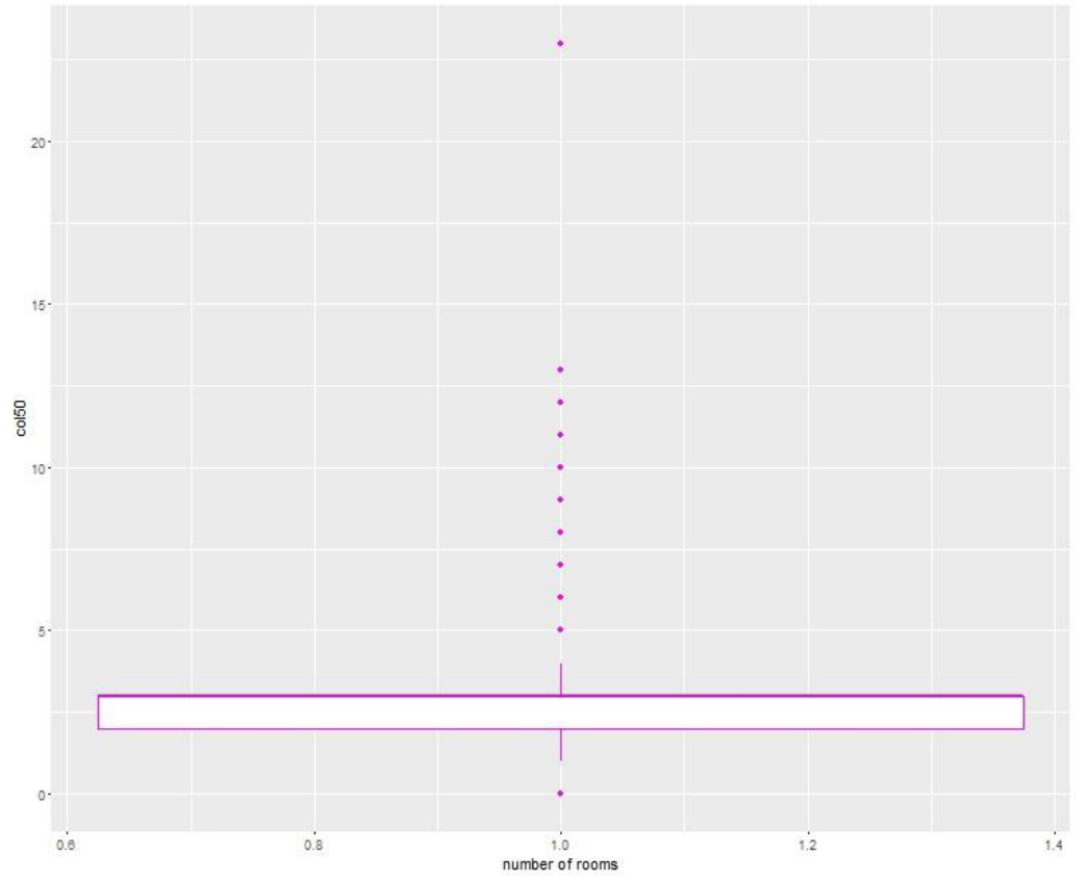
- تعداد اتاقهای در اختیار افراد بالای 25 سال



دادهها یونی مدال است در نگاه اول این دادهها کمی چوله به راست به نظرمیرسد و تعداد کمی مقدار خیلی بزرگ داریم ولی با چون میانه این دادهها 3 است و میانگین 2.92 و چون واریانس دادهها زیاد نیست میتوان گفت که اول چولگی کم است دوماً اگر طبق فرمول میانگین چولگی را حساب کنیم فرمول به ما اعلام میکند چولگی چپ داریم!!!!

$$\frac{3(mean - median)}{SD}$$

Q1

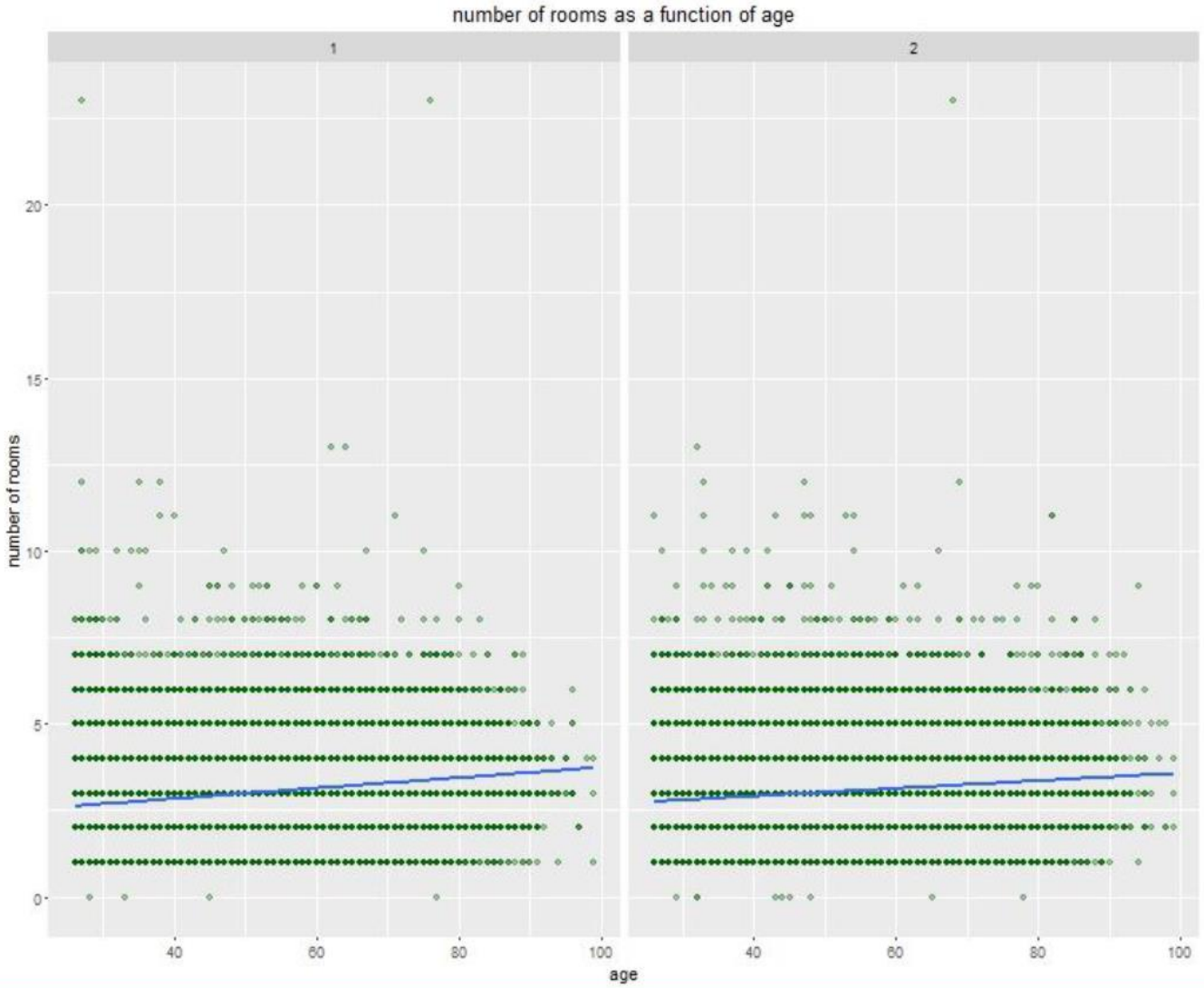


Q2

- دو متغیر عددی دلخواه انتخاب کنید و رابطه‌ی بین آنها را در قالب یک scatterplot نمایش دهید سپس این نمودار را تحلیل کنید. کورولویشن و کورایانس این دو متغیر را بدست آورید.

Q2

در این نمودار 1 برای مردان است و 2 برای زنان. مطمئناً بررسی دقیقتر لازم است که تفاوت معنا دار این دو مشخص شود ولی به نظر میرسد در این نمونه در سنین کم تعداد اتاقها برای زنان بیشتر است ولی شیب افزایش برای مردان بیشتر است.

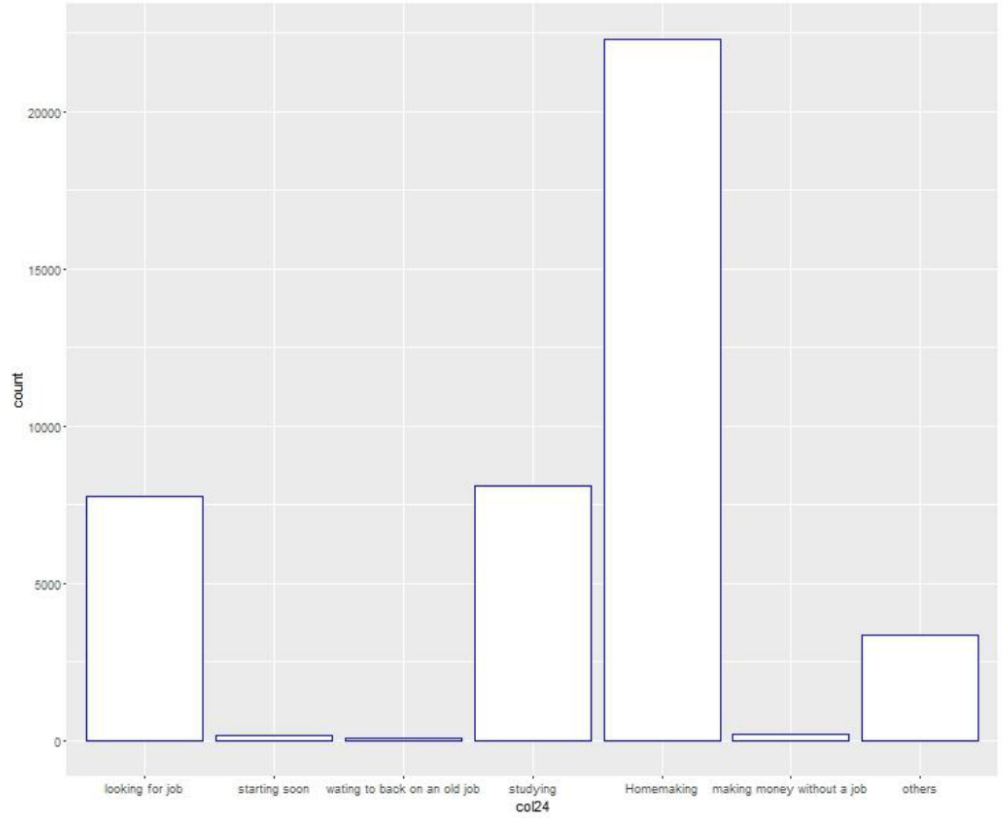


Q3

- یک متغیر رسته‌ای (Categorical) را به دلخواه انتخاب کرده و جدول فرکانسی و نمودار میله‌ای آن را ترسیم کنید.

Q3

- در این قسمت به وضعیت جستجوی کار در استان تهران در افراد بین 20 تا 35 سال نگاه میکنیم.



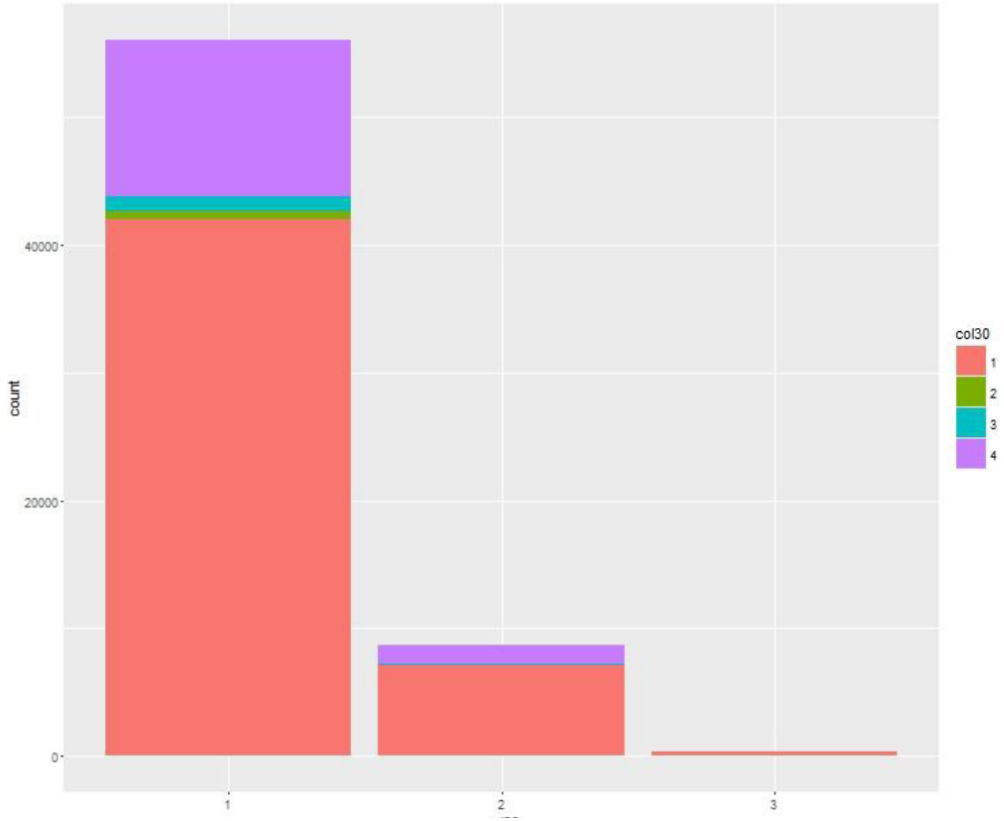
conditions	Looking for job	Starting soon	Waiting to back on an old job	Studying	Homemaking	Making money without a job	Others
Frequency	7750	186	82	8090	22317	212	3339

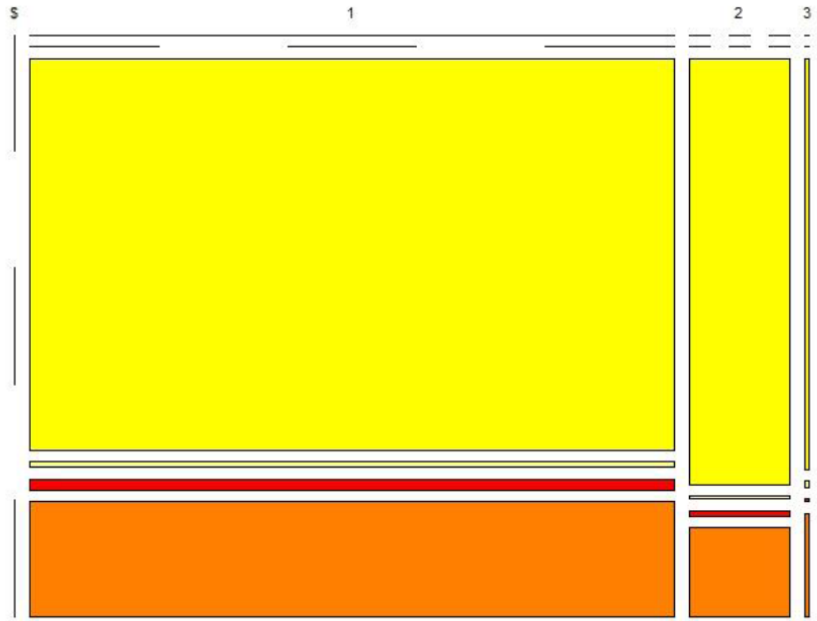
Q4

- دو متغیر رسته‌ای انتخاب کنید و برای این دو Contingency Table، Segmented Bar Plot و Mosaic Plot را نمایش دهید.

Q4

- در این قسمت میخواهیم ارتباط وضعیت شغلی و وضعیت تاهل را در استان تهران بررسی کنیم.





Q5

- یه متغیر عددی انتخاب کنید و برای میانگین آن بازه اطمینان 99 درصد را محاسبه کنید. سپس آن را تحلیل کنید.

Central Limit Theorem

Central Limit Theorem (CLT): The distribution of the sample mean is well approximated by a normal model:

$$\bar{x} \sim N(mean = \mu\ , SE = \frac{\sigma}{\sqrt{n}})$$

where SE represents standard error, which is defined as the standard deviation of the sampling distribution.

- Note that as n increases SE decreases.
- If σ is unknown, use s (the sample standard deviation).
  - s : the standard deviation of one sample that we happen to have at hand

Confidence Interval

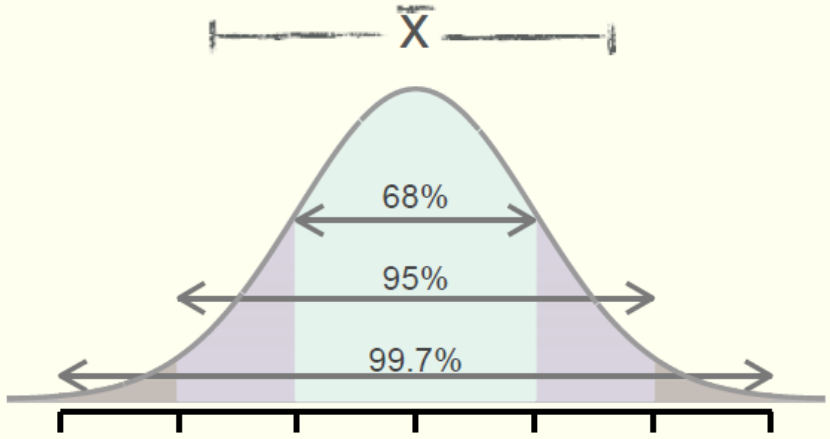
- A plausible range of values for the population parameter is called a confidence interval.

- If we report a point estimate, we probably won't hit the exact population parameter.
- If we report a range of plausible values we have a good shot at capturing the parameter.

Confidence Interval for $\bar{x}$

Central Limit Theorem (CLT)

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

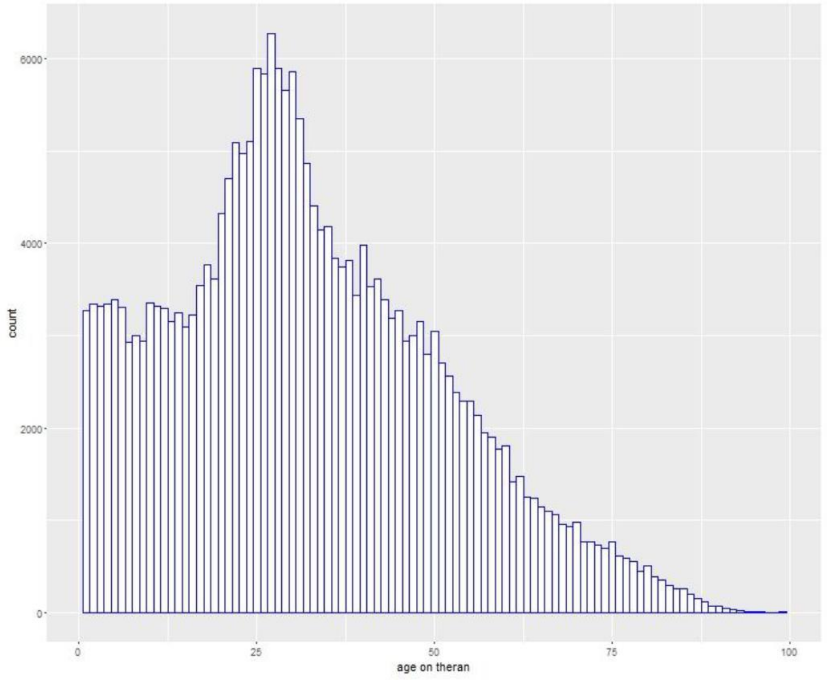


approximate 95% CI: $\bar{x} \pm 2SE$

margin of error (ME)

Q5

- در این قسمت میخواهیم میانگین سن افراد در استان تهران را بررسی کنیم.



$$CI = \bar{x} \pm t^{*}SE \rightarrow 32.32 \pm 0.099 \rightarrow [32.22,32.42]$$

```
ME <- qt(0.995,df=length(samp_a_age_tehran$Col07)-
1)*sd(samp_a_age_tehran$Col07)/sqrt(length(samp_a_age_tehran$Col07))

left <- mean(samp_a_age_tehran$Col07) - ME
right <- mean(samp_a_age_tehran$Col07) + ME
```

Q6

- برای میانگین یک متغیر عددی، آزمون فرضی را مطرح کنید و با محاسبه p-value فرض خود را تایید و یا رد کنید.

P-value

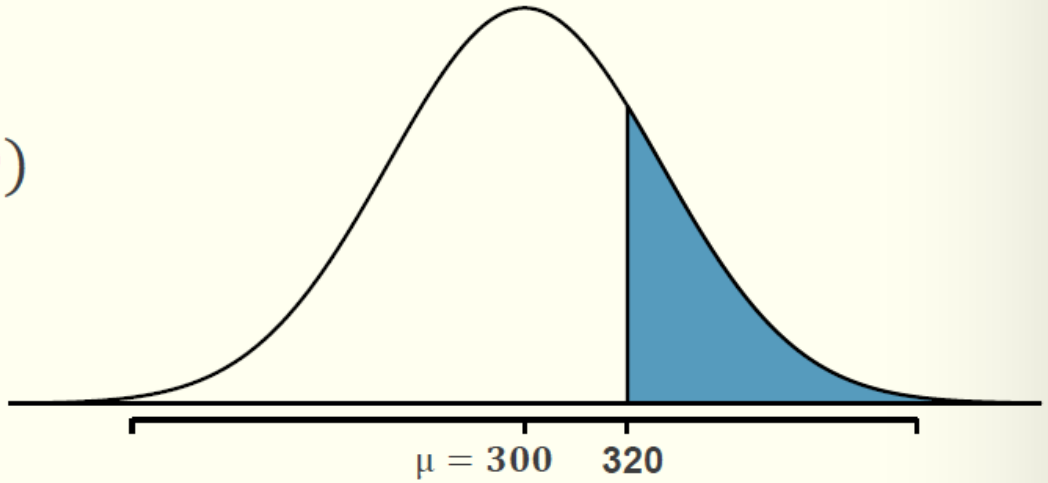
- p-value = P(observed or more extreme outcome | H_0 true)

For previous example:

p-value = $P(\bar{x} > 320 | H_0: \mu = 300)$

$s = 174\,, n = 50 \;\Rightarrow\; SE = 24.6$

$\bar{x} \sim N(\mu = 300\,, SE = 24.6)$



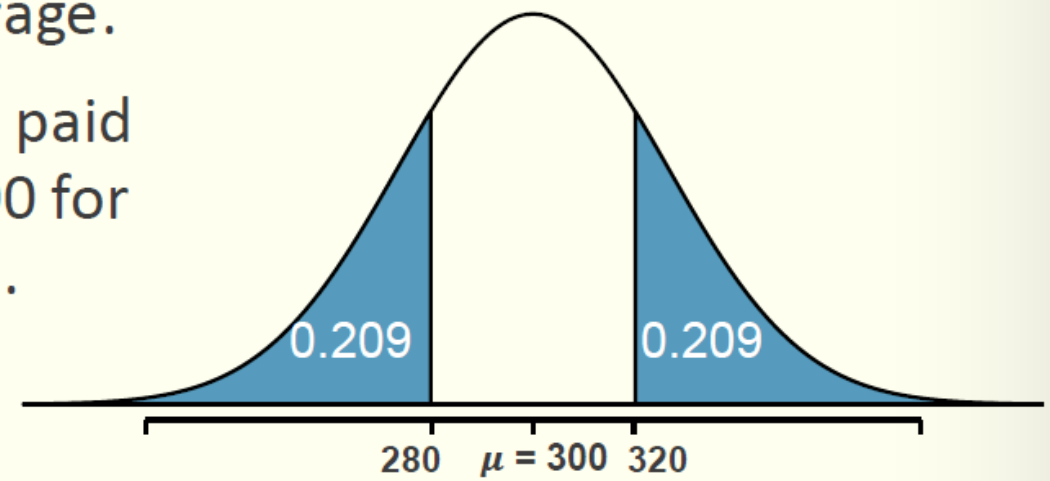
test statistics: $Z = \frac{320-300}{24.6} = 0.81$

p-value = $P(Z > 0.81) = 0.209$

Two-tailed p-value

$H_0: \mu = 300$ College students have paid \$300 for textbooks, on average.

$H_A: \mu \neq 300$ College students have paid more or less than \$300 for textbooks, on average.



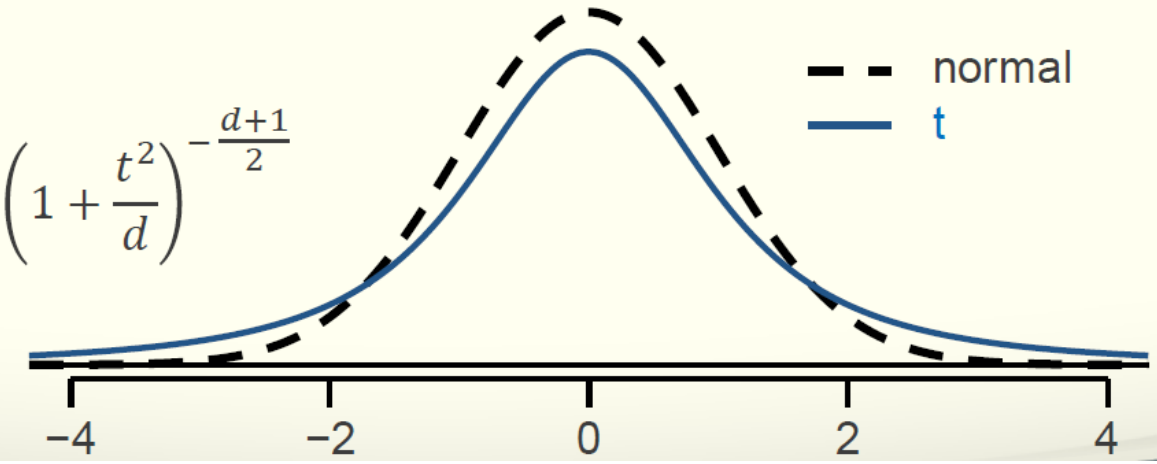
p-value = $P(\bar{x} > 320 \text{ or } \bar{x} < 280 | H_0: \mu = 300)$

p-value = $P(Z > 0.81) + P(Z < -0.81) = 0.209 + 0.209 = 0.418$

Student's t-distribution

- When σ is unknown (which is almost always), use the t-distribution to address the uncertainty of the standard error estimate
- Bell shaped but thicker tails than the normal
  - Observations more likely to fall beyond 2 SDs from the mean
  - Extra thick tails helpful for mitigating the effect of a less reliable estimate for the standard error of the sampling distribution

$$f(t) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{\pi d}\,\Gamma\left(\frac{d}{2}\right)}\left(1+\frac{t^2}{d}\right)^{-\frac{d+1}{2}}$$



t-statistic

- t-statistic is used for inference on a mean where:
  - σ unknown, which is almost always
- It is calculated the same way as z-statistic:

$$T = \frac{observation - null}{SE}$$

- p-value has also the same definition:

Example

- Find the following probabilities:

a. $P(|Z| > 2)$ 0.0455 → reject

b. $P(|t_{df=50}| > 2)$ 0.0509 → fail to reject?

c. $P(|t_{df=10}| > 2)$ 0.0734 → fail to reject

R

```
> pnorm(2, lower.tail = FALSE) * 2
  [1] 0.0455
> pt(2, df = 50, lower.tail = FALSE) * 2
  [1] 0.0509
```

Estimating the Mean

- point estimate ± margin of error

$$\bar{x} \pm t^{\star}_{df} SE_{\bar{x}}$$

$$\bar{x} \pm t^{\star}_{df} \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm t^{\star}_{n-1} \frac{s}{\sqrt{n}}$$

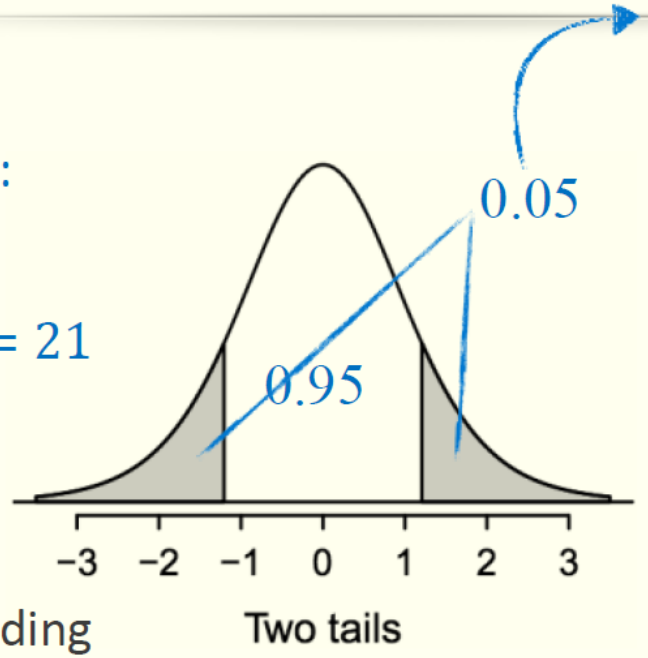
- Degrees of freedom for t statistic for inference on one sample mean:

$$df = n - 1$$

Finding the Critical t-score

a) Using the table:

1. determine df:
   $df = 22 - 1 = 21$



2. Find corresponding tail area for desired confidence level

	one tail	0.100	0.050	0.025	0.010	0.005
	two tails	0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17
	11	1.36	1.80	2.20	2.72	3.11
	12	1.36	1.78	2.18	2.68	3.05
	13	1.35	1.77	2.16	2.65	3.01
	14	1.35	1.76	2.14	2.62	2.98
	15	1.34	1.75	2.13	2.60	2.95
	16	1.34	1.75	2.12	2.58	2.92
	17	1.33	1.74	2.11	2.57	2.90
	18	1.33	1.73	2.10	2.55	2.88
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81
	24	1.32	1.71	2.06	2.49	2.80
	25	1.32	1.71	2.06	2.49	2.79
	26	1.31	1.71	2.06	2.48	2.78
	27	1.31	1.70	2.05	2.47	2.77

Example

- Estimate the average after-lunch snack consumption (in grams) of people who eat lunch **distracted** using a 95% confidence interval.

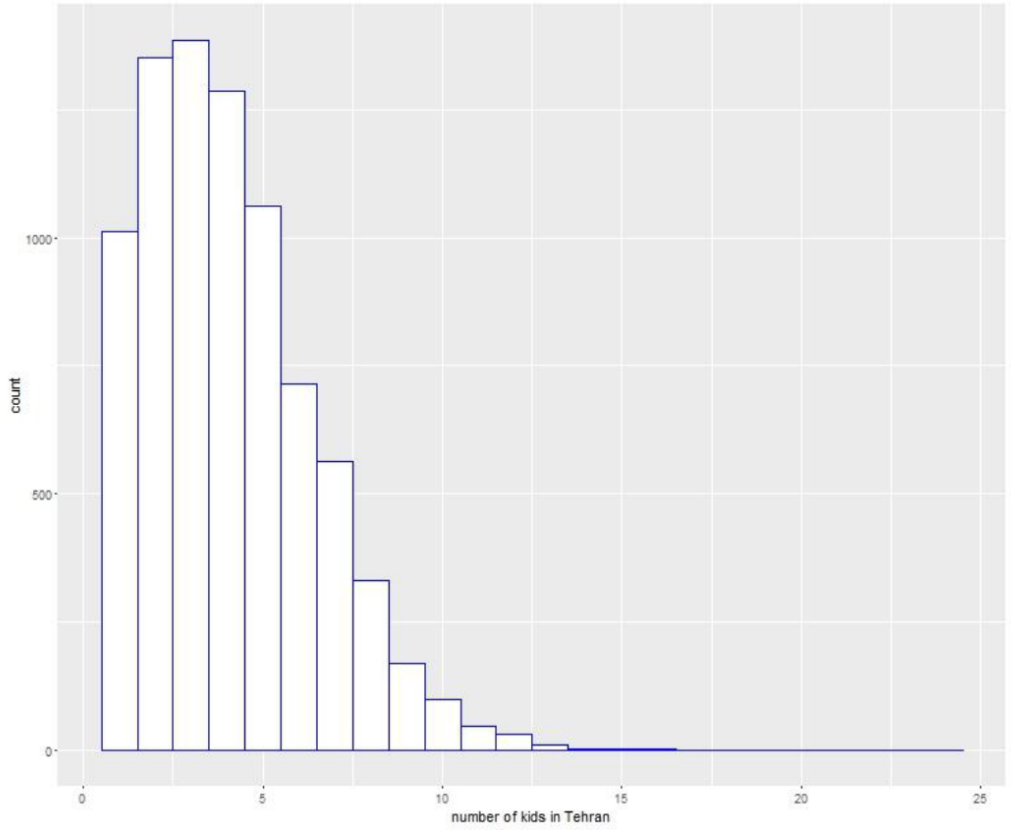
$$\bar{x} = 52.1\ g$$
$$s = 45.1\ g$$
$$n = 22$$
$$t^{\star}_{21} = 2.08$$

$$\bar{x} \pm t^{*}SE = 52.1 \pm 2.08 \times \frac{45.1}{\sqrt{22}}$$
$$= 52.1 \pm 2.08 \times 9.62$$
$$= 52.1 \pm 20 = (32.1, 72.1)$$

- We are 95% confident that distracted eaters consume between 32.1 to 72.1 grams of snacks post-meal.

Q6

- در این قسمت میخواهیم بررسی کنیم میانگین تعداد فرزندان به دنیا آمده در تهران چقدر است



آیا می‌توان فرض تعداد بچه‌ها در تهران ۴ است را در مقابل تعداد بچه‌ها بیشتر از ۴ است رد کرد

$H0 : \mu = 4$ (تعداد فرزندان به دنیا آمده سرپرست خانوارهای تهران برابر چهار است)

$HA : \mu > 4$ (تعداد فرزندان به دنیا آمده سرپرست خانوارهای تهران بیشتر چهار است)

```
t.test(samp_a_kids_tehran$all_kids, mu=3, alternative="less")
```

t = 2.9691, df = 8075, p-value = 0.001498

Q7

- یک متغیر پاسخ (response) عددی و تعدادی متغیر توضیحی عددی و رسته ای انتخاب کرده و معادله خط رگرسیون را محاسبه کنید. شیب همه متغیرهای توضیحی و عرض از مبدا را تفسیر کنید. برای حداقل یکی ازمتغیرهای توضیحی آزمون فرضی طراحی و اجرا کنید و نتیجه بگیرید آیا این متغیر توضیحی برای پیش بینی متغیر پاسخ مناسب است یا خیر.

Q7

- در این قسمت میخواهیم باز به سراغ میانگین تعداد فرزندان به دنیا آمده در تهران می‌رویم.

متغیرهای انتخاب شده برای تفسیر تعداد فرزندان به صورت زیر است:سن، آیا فرد ماشین دارد، با چند خانواده دیگر در یک خانه زندگی می‌کند، تعداد اتاقی که در خانه دارد، آیا بیشتر از یک محل سکونت دارد، آیا فرد در منزل به کامپیوتر دستری دارد

```
lm(all_kids ~ Col07 + col50 + col453+ col452+ col51+col09,
data = samp_a_kids_tehran)
```

تنها ضریبی که p-value بالایی دارد مربوط به متغیر چند خانه بودن است.

تفسیر ضرایب بسیار راحت است برای متغیرهای خطی مثل سن می‌گوییم به اعضای هر سال افزاش سن تعداد فرزند به اندازه ۰.۱۱۴ افزایش می‌یابد برای متغیرهای رسته-ای مثل حالت داشتن کامپیوتر را در نظر می‌گیریم و می‌گوییم اگر همه چیز مساوی باشد کسی که کامپیوتر ندارد به صورت میانگین ۰.۹۷ فرزند بیشتر دارد.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.920752   0.271097 -10.774  < 2e-16 ***
Col07        0.114608   0.002036  56.298  < 2e-16 ***
col50        0.124767   0.021045   5.929 3.21e-09 ***
col4532      0.978632   0.074757  13.091  < 2e-16 ***
col4522      0.409687   0.079745   5.137 2.86e-07 ***
col512       0.711757   0.090918   7.829 5.70e-15 ***
col092      -0.153918   0.389913  -0.395    0.693
col093       0.676884   0.431793   1.568    0.117
col094      -0.024860   0.241422  -0.103    0.918
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.383 on 6681 degrees of freedom
Multiple R-squared:  0.4151,   Adjusted R-squared:  0.4144
F-statistic: 592.8 on 8 and 6681 DF,  p-value: < 2.2e-16
```

عرض از مبدا در اینجا مربوط به کسی می‌شود که سن صفر دارد که از آنجا که کسی که سن صفر ندارد مطمئن بچه ندارد این فقط برای تنظیم نمودار است و معنی خاصی ندارد.