امتحان پایان‌ترم استنباط آماری

۱. **(۲۰ نمره)** فروشگاهی به دنبال بررسی اثر دو روش تبلیغاتی مختلف و قیمت یک محصول بر میزان فروش آن محصول است. مجموعه‌داده‌ای به صورت هفتگی توسط این فروشگاه جمع‌آوری شده است که هر سطر آن متعلق به یک هفته است و از چهار متغیر sales (تعداد واحد فروش محصول)، price (قیمت محصول در آن هفته)، A1 (متغیر باینری که مقدار آن 1 است اگر تبلیغ نوع اول در آن هفته استفاده شده باشد)، و A2 (متغیر باینری که مقدار آن 1 است اگر تبلیغ نوع دوم در آن هفته استفاده شده باشد).

در ابتدا برای تحلیل از یک مدل رگرسیون خطی ساده استفاده شده است و خط رگرسیون به صورت

$$\text{sales} = 2258 - 1417 \times \text{price}$$

به دست آمده است. مقدار p-value تست ANOVA برای این رگرسیون برابر 0.000 و $R^2 = 0.596$ است.

الف) شیب خط رگرسیون را تفسیر کنید. (۱)

As the price increases by 1 dollar, sales will decrease, on average, by 1417 units.

ب) آیا عرض از مبدا خط رگرسیون تفسیرپذیر است؟ چرا؟ (۱)

No, since a price of zero dollars is probably out of the range observed.

پ) ضریب همبستگی خطی بین دو متغیر sales و price چقدر است؟ (۱)

$$R^2 = 0.596 \rightarrow \ |R| = \sqrt{0.596} = 0.772 \ \text{, negative slope} \ \rightarrow R = -0.772$$

ت) آیا متغیر price پیش‌بینی‌کننده خوبی برای sales است؟ چرا؟ (۱)

Yes, the p-value of the model is very small.

ث) اگر residual plot این تحلیل تقریبا به صورت شکل زیر باشد، چه نتیجه‌ای راجع به تحلیل انجام شده می‌توان گرفت؟ (۱)

The assumption of constant variability might be violated.

در مرحله بعد متغیر price2 که در واقع مربع متغیر price است، به مدل اضافه می‌شود. خروجی مدل به صورت زیر است:

Predictor	Coef	s.d.	t-statistic	p-value
Constant	7990.0	724.7	11.03	0.000
Price	-10660	1151	-9.26	0.000
Price2	3522.3	436.8	8.064	0.000

Analysis of Variance:

Source	DF	SS	MS	F	p-value
Regression	2	16060569	8030284	125.11	0.000
Error	60	3851231	64187		
Total	62	19911800			

ج) جاهای خالی را در جداول بالا پر کنید. محاسبات مربوط به تعیین هر مقدار را به طور کامل نمایش دهید. (۳)

چ) مقدار R^2 در این مدل چقدر است؟ (۱)

$$R^2 = \frac{16060569}{19911800} = 0.8066$$

ح) این مجموعه داده چند سطر دارد؟ (۱)

$$n - 1 = 62 \rightarrow \ n = 63$$

خ) آیا این مدل از رگرسیون خطی قبلی بهتر است؟ چرا؟ (۲)

In model 1: $R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-k-1}\right) = 1 - \left((1 - 0.596) \times \frac{62}{61}\right) = 0.59$

In model 2: $R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-k-1}\right) = 1 - \left(\frac{3851231}{19911800} \times \frac{62}{60}\right) = 0.8$

Thus model 2 has more predictive power and is better than model 1.

د) برای این مدل residual plot به چه شکلی خواهد بود؟ (۱)

It will have better constant variability.

در مرحله بعد متغیرهای باینری A1 و A2 هم به مدل اضافه می‌شوند. خروجی مدل به صورت زیر است:

Predictor	Coef	s.d.	t-statistic	p-value
Constant	3829.5	700.4	5.47	0.000
Price	-5056	1026	-4.93	0.000
Price2	1667.7	369.9	4.51	0.000
A1	804.12	86.75	9.27	0.000
A2	-31.49	53.38	-0.59	0.558

Analysis of Variance:

Source	DF	SS	MS	F	p-value
Regression	4	18373664	4593416	173.21	0.000
Error	58	1538137	26520		
Total	62	19911802			

ذ) جداول بالا را تکمیل کنید. محاسبات مربوط به تعیین هر مقدار را به طور کامل نمایش دهید. (۳)

ر) مقدار R^2_{adj} را حساب کنید. (۱)

$$R^2_{adj} = 1 - \left(\frac{1538137}{19911802} \times \frac{62}{58}\right) = 0.9174$$

ز) آیا روش‌های تبلیغی به کار گرفته شده توسط فروشگاه موثرند؟ چرا؟ (۱)

A1 is effective (small p-value in the model) but A2 is ineffective (large p-value).

ژ) آیا می‌توان این مدل را بهبود بخشید؟ توضیح دهید. (۱)

.Yes. A2 is not a good predictor and should be removed

س) می‌دانیم تبلیغ نوع اول معمولا در زمانی به کار گرفته می‌شود که فروشگاه تصمیم دارد محصول مورد نظر را در طول یک هفته خاص با تخفیف به فروش برساند. آیا می‌توان این اطلاعات اضافی را در مدل بالا وارد کرد؟ توضیح دهید. (۱)

Yes, we can use a new variable which is equal to the product of A1 and price and models the .interaction between these two variables

۲. **(۵ نمره)** مدیر یک سینمای کوچک تک‌سالنی به دنبال این است که مشخص سازد آیا ارتباطی بین ژانر فیلم نمایش داده شده در سالن این سینما و مصرف تنقلات توسط تماشاچیان فیلم وجود وجود دارد یا خیر. او با بررسی تعداد تماشاچیان چهار ژانر مختلف از فیلم‌ها و مصرف تنقلات توسط آنها، مجموعه داده زیر را تهیه می‌کند:

ژانر فیلم / مصرف تنقلات	بله	خیر
اکشن	۵۰	۷۵
کمدی	۱۲۵	۱۷۵
خانوادگی	۹۰	۳۰
ترسناک	۴۵	۱۰

با طراحی و اجرای یک آزمون فرض مناسب به سوال مدیر سینما پاسخ دهید.

H_0 : snacking and genre are independent

H_A : snacking and genre are dependent

Conditions:

1. ***Independence:*** Sampled observations are independent: we have random sample/assignment, $n = 600 < 10\%$ of population, each case only contributes to one cell in the table
2. ***Sample size:*** Each cell has more than 5 expected cases.

$df = (R-1)(C-1) = 3$

No	**Yes**	
$75\ (\frac{125\times290}{600} = 60.42)$	$50\ (\frac{125\times310}{600} = 64.58)$	**125**
$175\ (\frac{300\times290}{600} = 145)$	$125\ (\frac{300\times310}{600} = 155)$	**300**
$30\ (\frac{120\times290}{600} = 58)$	$90\ (\frac{120\times310}{600} = 62)$	**120**
$10\ (\frac{55\times290}{600} = 26.58)$	$45\ (\frac{55\times310}{600} = 28.42)$	**55**
290	**310**	

$$\chi^2 = \frac{(75-60.42)^2}{60.42} + \frac{(50-64.58)^2}{64.58} + \frac{(175-145)^2}{145} + \frac{(125-155)^2}{155} + \frac{(30-58)^2}{58} + \frac{(90-62)^2}{62} + \frac{(10-26.58)^2}{26.58} + \frac{(45-28.42)^2}{28.42} = 65$$

p-value = $P(\chi_3^2 > 65) \approx 0$

p-value =0 $\rightarrow$ reject H_0 $\rightarrow$ snacking and genre are dependent

۳. **(۵ نمره)** یک استاد درس یادگیری ماشین می‌خواهد بررسی کند آیا از نظر آماری اختلاف معناداری بین نرخ مردودی دانشجویان دکترا و کارشناسی‌ارشدی که این درس را داشته‌اند وجود دارد یا خیر. او به صورت تصادفی ۱۰۷ نفر از دانشجویان دکترا و ۱۴۳ نفر از دانشجویان کارشناسی‌ارشدی که این درس را داشته‌اند انتخاب می‌کند. طبق مشاهدات او ۳۰ نفر از دانشجویان دکترا و ۴۵ نفر از دانشجویان کارشناسی‌ارشد در این درس رد شده‌اند. با طراحی و اجرای یک آزمون فرض مناسب به سوال مورد تحقیق این استاد پاسخ دهید.

H_0: $p_1 - p_2 = 0$

H_A: $p_1 - p_2 \neq 0$

$$\hat{p}_{pool} = \frac{30 + 45}{107 + 143} = \frac{75}{250} = 0.3$$

Condition:

$$n_1\hat{p}_{pool} = 107 \times 0.3 > 10$$

$$n_1(1 - \hat{p}_{pool}) = 107 \times 0.7 > 10$$

$$n_2\hat{p}_{pool} = 143 \times 0.3 > 10$$

$$n_2(1 - \hat{p}_{pool}) = 143 \times 0.7 > 10$$

$$SE = \sqrt{\frac{\hat{p}_{pool}(1 - \hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1 - \hat{p}_{pool})}{n_2}} = \sqrt{\frac{0.3 \times 0.7}{107} + \frac{0.3 \times 0.7}{143}} = 0.0586$$

point estimate $= \hat{p}_1 - \hat{p}_2 = \frac{30}{107} - \frac{45}{143} = -0.0343$

$Z = \frac{-0.0343 - 0}{0.0586} \approx -0.585 \rightarrow$ p-value = 2*P(Z < -0.585) = 0.56 > 0.05 $\rightarrow$ We fail to reject H_0

۴. **(۸ نمره)** برای هر یک از گزاره‌های زیر مشخص کنید آیا مغالطه آماری وجود دارد یا خیر. پاسخ خود را به طور کامل توضیح داده و در صورت وجود مغالطه آماری نحوه تصحیح آن را شرح دهید.

الف) به گفته معاون امور جوانان و ساماندهی وزارت ورزش و جوانان آمار طلاق در سال ۱۳۹۶ نسبت به سال قبل از آن از ۱۸۱۰۴۹ مورد به ۱۷۴۵۹۰ مورد کاهش یافته است که این کاهش ۳/۵۷ درصدی نشان از عملکرد موفق این سازمان دارد.

کاهش ۳/۵ درصدی طلاق به کاهش ۱۳ درصدی ازدواج مرتبط است و به عملکرد این سازمان ربطی ندارد.

ب) طبق اعلام مرکز ملی آمار افزایش متوسط درآمد سالانه خانوار شهری و روستایی در سال ۱۳۹۹ نسبت به سال ۱۳۹۸ به ترتیب ۲۴/۴ و ۲۷/۴ درصد بوده است در حالی که افزایش متوسط هزینه سالانه خانوار شهری و روستایی در مدت زمان مشابه به ترتیب ۲۰/۶ و ۲۱/۷ درصد اعلام شده، به عبارت دیگر با وجود تورم اوضاع معیشتی مردم رو به بهبود است.

میانگین آماره مناسبی نیست. تورم موجب افزایش اختلاف طبقاتی شده و حتی در صورت افزایش میانگین لزوما وضعیت معیشتی اکثریت جامعه بهبود نیافته است.

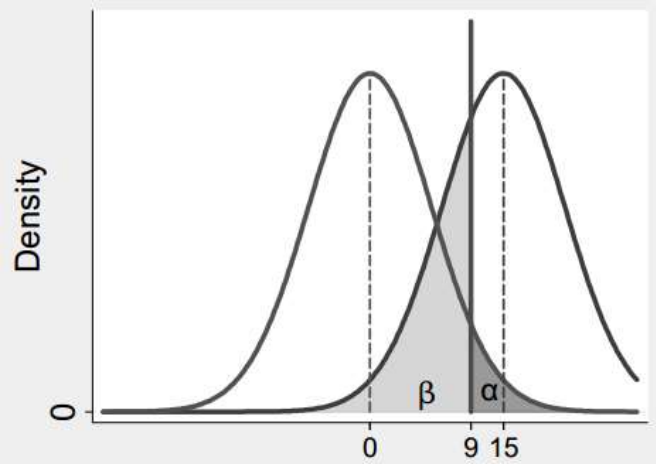
پ) افزایش میزان باسوادی در کشور از ۸۴/۶ درصد در سال ۱۳۸۵ به ۸۷/۶ درصد در سال ۱۳۹۵، با افزایش نرخ بیکاری از ۱۰/۳ درصد به ۱۳/۵ درصد همراه بوده است، بنابراین بر خلاف نظریه‌های جامعه‌شناسان غربی، سوادآموزی موجب از میان رفتن بیکاری نمی‌شود.

مغالطه علت شمردن همبستگی اتفاق افتاده است. بیکاری به عوامل متعدد بستگی دارد و صرف همبستگی بالای این دو متغیر به معنای این نیست که یکی علت دیگری است.

ت) برای بررسی نظر دانشجویان درباره کیفیت غذای سلف، کافی است در جلوی درب خروجی سلف بایستیم و از هر ۵ نفری که از سلف خارج می‌شوند یک نفر را به تصادف انتخاب کرده و در مورد کیفیت غذا از او سوال کنیم.

False. This approach for sampling suffers from convenience bias.

۵. **(۶ نمره)** نمودار زیر خطای نوع اول و دوم را به صورت گرافیکی برای یک آزمون فرض در ارتباط با میانگین یک جامعه آماری که با استفاده از نمونه‌ای با اندازه ۱۰۰ صورت گرفته نمایش می‌دهد:



الف) آزمون فرض مرتبط با این نمودار را مشخص کنید.

ب) با فرض $\alpha = 0.05$ توان این آزمون را بیابید.

(a) $H_0: \mu = 0$
$H_A: \mu > 0$

(b) $P(\overline{X_1} > 9) = P\left(Z > \frac{9-0}{SE}\right) = 0.05 \rightarrow \frac{9}{SE} = 1.645 \rightarrow SE = 5.47$

$$power = 1 - \beta = P(\overline{X_2} > 9) = P\left(Z > \frac{9-15}{SE}\right) = P(Z > -1.1)$$
$$= P(Z < 1.1) = 0.8643$$

۶. **(۶ نمره)** یک درصد از مشتریان یک بانک اقساط وام خود را پرداخت نمی‌کنند. این بانک مشتریان خود را به سه دسته با ریسک کم، متوسط، و زیاد تقسیم می‌کند. ۳۰٪ از وام‌های پرداخت‌نشده به مشتریان کم‌ریسک، ۴۰٪ به مشتریان با ریسک متوسط، و ۳۰٪ به مشتریان با ریسک بالا تعلق دارند. از وام‌های پرداخت‌شده ۵۰٪ به مشتریان کم‌ریسک، ۴۰٪ به مشتریان با ریسک متوسط، و ۱۰٪ به مشتریان با ریسک بالا تعلق دارند.

الف) احتمال این که یک مشتری با ریسک بالا، وام خود را پرداخت نکند چقدر است؟

ب) احتمال این که یک مشتری با ریسک پایین وام خود را پرداخت کند چقدر است؟

الف) اگر A پیشامد پرداخت وام و A' پیشامد پرداخت نکردن وام باشند:

L: Low risk

$$P(L|A') = 0.3 \ , \ P(L|A) = 0.5$$

M: Medium risk

$$P(M|A') = 0.4 \ , \ P(M|A) = 0.4$$

H: High risk

$$P(H|A') = 0.3 \ , \ P(H|A) = 0.1$$

$$P(A'|H) = \frac{P(H|A')P(A')}{P(H|A')P(A') + P(H|A)P(A)} = \frac{0.3 \times 0.01}{0.3 \times 0.01 + 0.1 \times 0.99} = 0.03$$

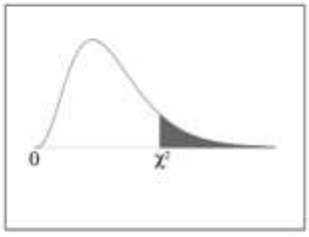
ب)

$$P(A|L) = \frac{P(L|A)P(A)}{P(L|A)P(A) + P(L|A')P(A')} = \frac{0.5 \times 0.99}{0.5 \times 0.99 + 0.3 \times 0.01} = 0.994$$

Standard Normal Table

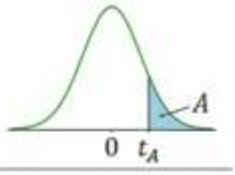
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750



Critical Values of t:

ν	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$	ν	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.657	38	1.304	1.686	2.024	2.429	2.712
2	1.886	2.920	4.303	6.965	9.925	39	1.304	1.685	2.023	2.426	2.708
3	1.638	2.353	3.182	4.541	5.841	40	1.303	1.684	2.021	2.423	2.704
4	1.533	2.132	2.776	3.747	4.604	41	1.303	1.683	2.020	2.421	2.701
5	1.476	2.015	2.571	3.365	4.032	42	1.302	1.682	2.018	2.418	2.698
6	1.440	1.943	2.447	3.143	3.707	43	1.302	1.681	2.017	2.416	2.695
7	1.415	1.895	2.365	2.998	3.499	44	1.301	1.680	2.015	2.414	2.692
8	1.397	1.860	2.306	2.896	3.355	45	1.301	1.679	2.014	2.412	2.690
9	1.383	1.833	2.262	2.821	3.250	46	1.300	1.679	2.013	2.410	2.687
10	1.372	1.812	2.228	2.764	3.169	47	1.300	1.678	2.012	2.408	2.685
11	1.363	1.796	2.201	2.718	3.106	48	1.299	1.677	2.011	2.407	2.682
12	1.356	1.782	2.179	2.681	3.055	49	1.299	1.677	2.010	2.405	2.680
13	1.350	1.771	2.160	2.650	3.012	50	1.299	1.676	2.009	2.403	2.678
14	1.345	1.761	2.145	2.624	2.977	51	1.298	1.675	2.008	2.402	2.676
15	1.341	1.753	2.131	2.602	2.947	52	1.298	1.675	2.007	2.400	2.674
16	1.337	1.746	2.120	2.583	2.921	53	1.298	1.674	2.006	2.399	2.672
17	1.333	1.740	2.110	2.567	2.898	54	1.297	1.674	2.005	2.397	2.670
18	1.330	1.734	2.101	2.552	2.878	55	1.297	1.673	2.004	2.396	2.668
19	1.328	1.729	2.093	2.539	2.861	60	1.296	1.671	2.000	2.390	2.660
20	1.325	1.725	2.086	2.528	2.845	65	1.295	1.669	1.997	2.385	2.654
21	1.323	1.721	2.080	2.518	2.831	70	1.294	1.667	1.994	2.381	2.648
22	1.321	1.717	2.074	2.508	2.819	75	1.293	1.665	1.992	2.377	2.643
23	1.319	1.714	2.069	2.500	2.807	80	1.292	1.664	1.990	2.374	2.639
24	1.318	1.711	2.064	2.492	2.797	90	1.291	1.662	1.987	2.368	2.632
25	1.316	1.708	2.060	2.485	2.787	100	1.290	1.660	1.984	2.364	2.626
26	1.315	1.706	2.056	2.479	2.779	120	1.289	1.658	1.980	2.358	2.617
27	1.314	1.703	2.052	2.473	2.771	140	1.288	1.656	1.977	2.353	2.611
28	1.313	1.701	2.048	2.467	2.763	160	1.287	1.654	1.975	2.350	2.607
29	1.311	1.699	2.045	2.462	2.756	180	1.286	1.653	1.973	2.347	2.603
30	1.310	1.697	2.042	2.457	2.750	200	1.286	1.653	1.972	2.345	2.601
31	1.309	1.696	2.040	2.453	2.744	250	1.285	1.651	1.969	2.341	2.596
32	1.309	1.694	2.037	2.449	2.738	300	1.284	1.650	1.968	2.339	2.592
33	1.308	1.692	2.035	2.445	2.733	400	1.284	1.649	1.966	2.336	2.588
34	1.307	1.691	2.032	2.441	2.728	500	1.283	1.648	1.965	2.334	2.586
35	1.306	1.690	2.030	2.438	2.724	750	1.283	1.647	1.963	2.331	2.582
36	1.306	1.688	2.028	2.434	2.719	1000	1.282	1.646	1.962	2.330	2.581
37	1.305	1.687	2.026	2.431	2.715	∞	1.282	1.645	1.960	2.326	2.576

Degrees of freedom: ν