ضروریات آمار برای علم داده

آزمون فرض‌های آماری و مفهوم P مقدار

آزمون فرض

- یک فرض آماری ادعایی در مورد یک یا چند جمعیت مورد بررسی است که ممکن است درست یا نادرست باشد. به عبارت دیگر فرض آماری یک ادعا یا گزاره‌ای در مورد توزیع یک جمعیت یا توزیع یک متغیر تصادفی است.

- در آزمون فرض آماری به بررسی صحت ادعای انجام شده روی خصوصیتی از جامعه، بر اساس اطلاعات به دست آمده از داده‌های نمونه پرداخته می‌شود.

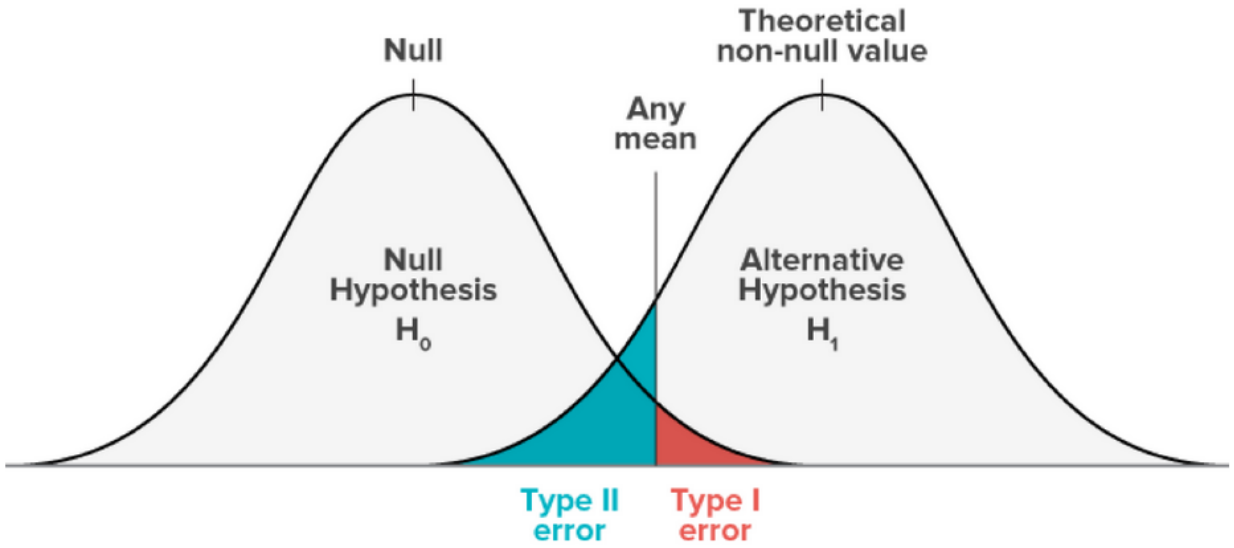
H_0 : ادعا درست است
H_1 : ادعا نادرست است

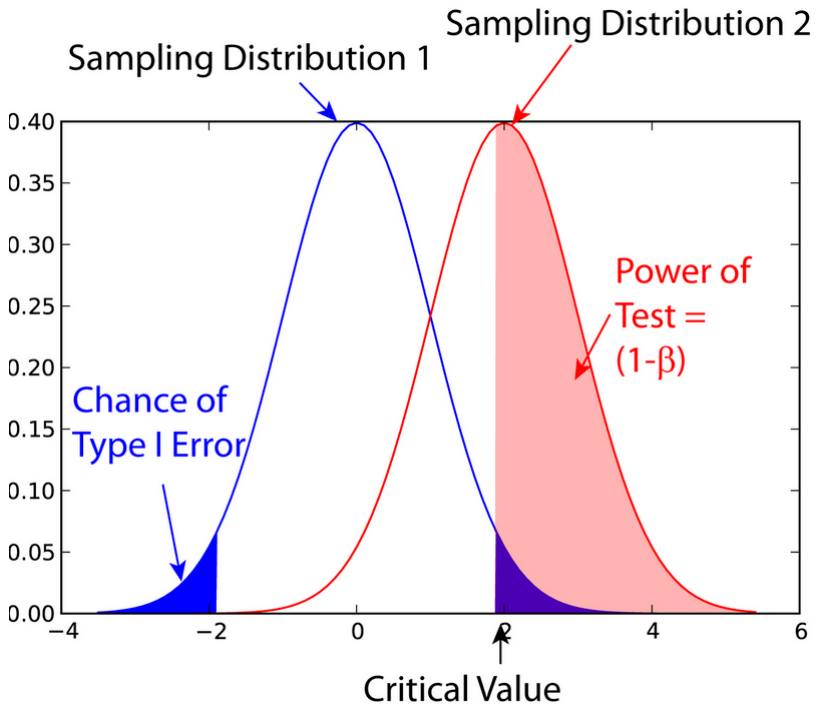
خطای نوع اول، دوم و توان آزمون

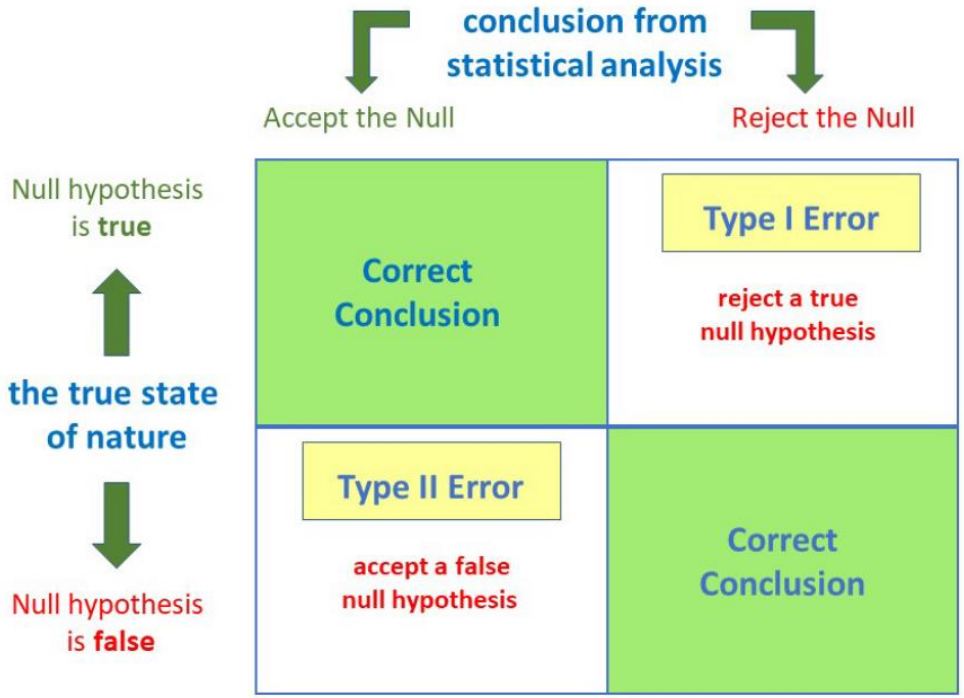
$\alpha = P(H_0$ رد | H_0 درست است)

$B = P(H_0$ پذیرش | H_1 درست است)

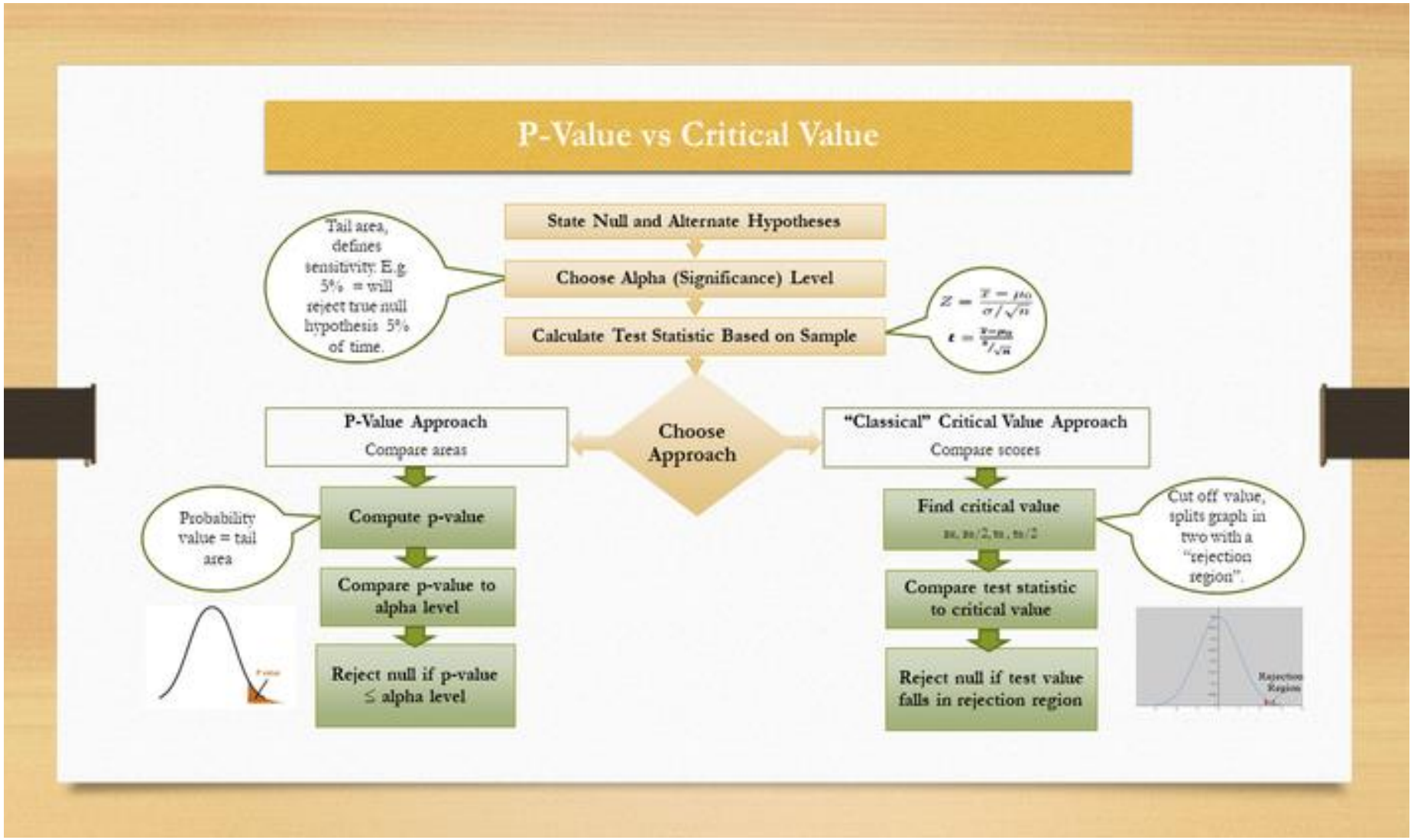
$B^* = 1 - B$






conclusion from
statistical analysis
Accept the Null
Reject the Null
Null hypothesis
is true
the true state
of nature
Null hypothesis
is false
Correct
Conclusion
Type I Error
reject a true
null hypothesis
Type II Error
accept a false
null hypothesis
Correct
Conclusion

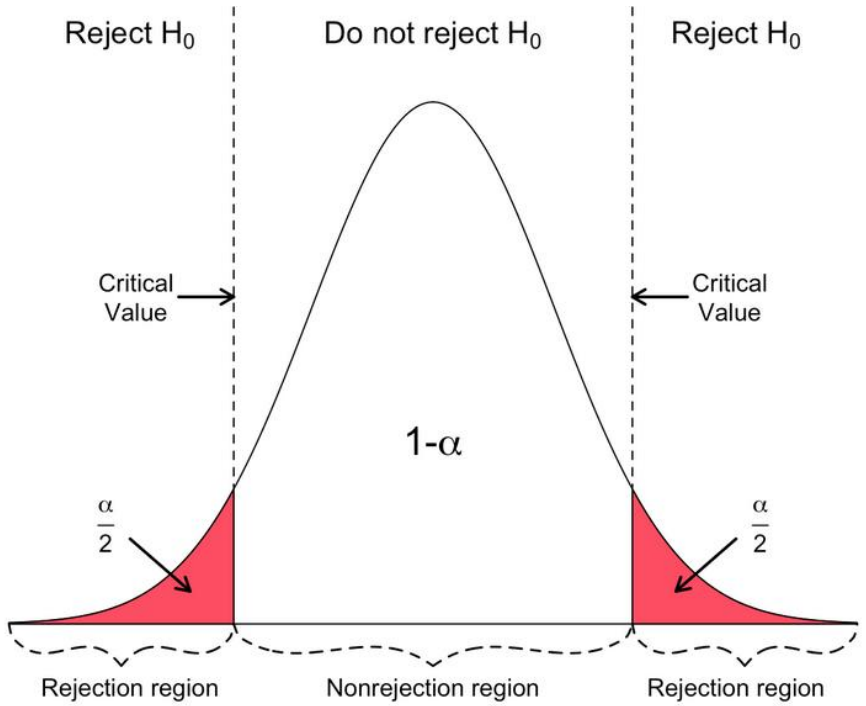
	Null hypothesis is true	**Null hypothesis is false**
Reject the null **hypothesis**	**Type I error** False positive	Correct outcome True positive
Fail to reject the **null hypothesis**	Correct outcome True negative	**Type II error** False negative


P-Value vs Critical Value
Tail area, defines sensitivity. E.g. 5% = will reject true null hypothesis 5% of time.
State Null and Alternate Hypotheses
Choose Alpha (Significance) Level
Calculate Test Statistic Based on Sample
Choose Approach
P-Value Approach
Compare areas
Probability value = tail area
Compute p-value
Compare p-value to alpha level
Reject null if p-value ≤ alpha level
"Classical" Critical Value Approach
Compare scores
Find critical value
Cut off value, splits graph in two with a "rejection region".
Compare test statistic to critical value
Reject null if test value falls in rejection region

سطح معنی داری

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 \end{cases}$$

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$



جامعه آماری، از پسرانی که در محدوده سنی ۱۰ تا ۱۲ سال هستند تشکیل شده است. داده‌های قبلی نشان می‌دهد که متوسط قد این افراد برابر است با ۷۵ سانتی‌متر و واریانس جامعه آماری برای قد این پسران برابر است با ۱۱.۶۲ سانتی‌متر مربع. با توجه به تغییر شیوه تغذیه اعتقاد داریم که میانگین قد پسرها در جامعه افزایش داشته و به ۸۰ سانتی‌متر رسیده است. براساس یک نمونه ۲۵ تایی میانگین قدها برابر با ۸۰.۹۴ سانتی‌متر بدست آمده است. آیا می‌توان از اطلاعات قبلی در مورد قد اطمینان داشت یا می‌توان به کمک آزمون آماری نشان داد که تغییر محسوسی در میزان قد پسران رخ داده است؟

$$\begin{cases} H_0 : \mu = 75 \\ H_1 : \mu = 80 \end{cases}$$

فرض صفر	فرض مقابل	آماره آزمون	ناحیه بحرانی
$\mu = \mu_0$	$\mu = \mu_1, \quad \mu_1 > \mu_0$	$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$	$Z > z_{(1-\alpha)}$
$\mu = \mu_0$	$\mu < \mu_0$	$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$	$Z < -z_{(1-\alpha)}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$	$\lvert Z \rvert > z_{(1-\frac{\alpha}{2})}$

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} =$$

$$Z = \frac{80.94 - 75}{\frac{11.6}{\sqrt{25}}} =$$

$$Z = \frac{5.94}{2.32} = 2.56$$

$$Z > z_{(1-\alpha)}$$

$$2.56 > 1.64$$

Standard Normal Probabilities

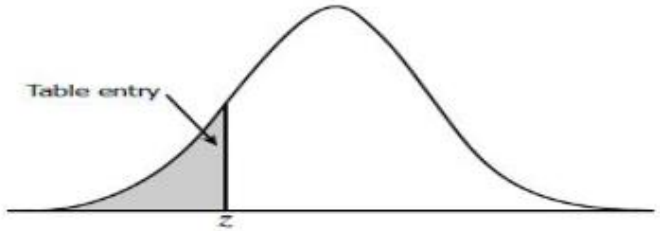


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Standard Normal Probabilities

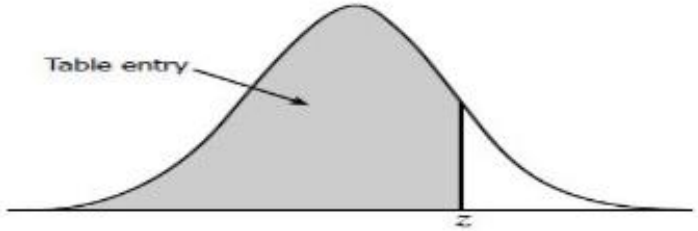


Table entry for z is the area under the standard normal curve to the left of z.

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
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

یک شرکت تولید کننده موتورسیکلت ادعا دارد که میزان مصرف سوخت تولیداتش در هر ۱۰۰ کیلومتر برابر با ۲ لیتر است. به این منظور سازمان بهینه‌سازی مصرف سوخت، ۸ موتورسیکلت از این شرکت را به منظور بررسی ادعایش تحویل گرفته. اطلاعات مربوط به مصرف سوخت این ۸ دستگاه در جدول زیر آورده شده است. در سطح خطای $\alpha=0.05$ ادعای تولید کننده بررسی می‌شود.

شماره نمونه	1	2	3	4	5	6	7	8
مصرف سوخت	3.0	2.8	3.2	2.6	3.3	2.5	2.8	3.0

$$\overline{X} = 2.9, \quad s(X) = 0.278, \quad n = 8$$

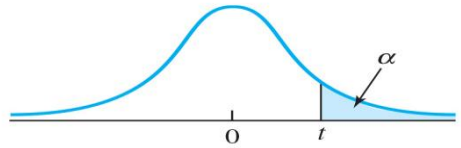
$$T = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} =$$

$$\frac{2.9 - 2}{\frac{0.278}{\sqrt{8}}} = 9.165$$

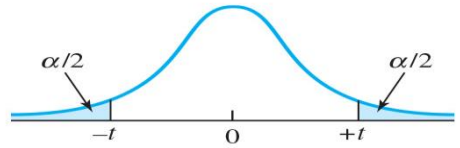
۱.۸۹۵<۹.۱۶۵ پس ادعای کارخانه تولید کننده موتورسیکلت رد می‌شود.

فرض صفر	فرض مقابل	آماره آزمون	ناحیه بحرانی
$\mu = \mu_0$	$\mu = \mu_1, \quad \mu_1 > \mu_0$	$T = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$	$T > t_{(1-\alpha)}(n-1)$
$\mu = \mu_0$	$\mu < \mu_0$	$T = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$	$T < -t_{(1-\alpha)}(n-1)$
$\mu = \mu_0$	$\mu \neq \mu_0$	$T = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$	$\lvert T \rvert > t_{(1-\frac{\alpha}{2})}(n-1)$

Percentage Points of the t Distribution



One-Tailed Test

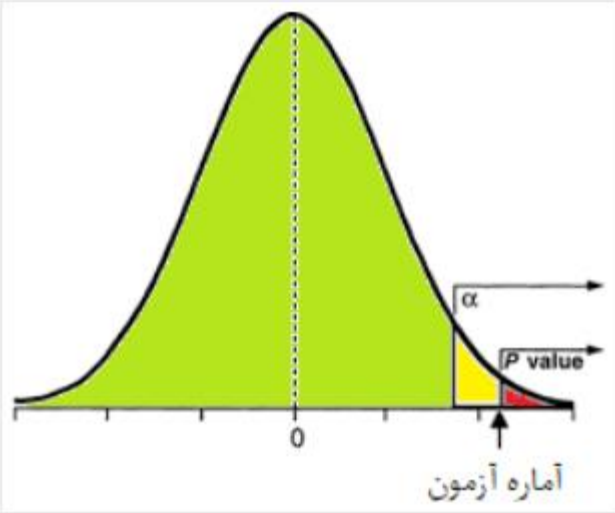


Two-Tailed Test

	Level of Significance for One-Tailed Test								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	Level of Significance for Two-Tailed Test								
df	.50	.40	.30	.20	.10	.05	.02	.01	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	63.662
2	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.403	2.678	3.496
100	.677	.845	1.042	1.290	1.660	1.984	2.364	2.626	3.390
∞	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

P- مقدار

- کمترین مقداری از خطای نوع اول (سطح آزمون) است، که یافته آماره آزمون ممکن است موجب رد فرض صفر شود. به بیان دیگر، در یک آزمون فرض، p-value برابر با کمترین مقداری از سطح معنی‌داری (significance level) یا همان احتمال خطای نوع اول است، که موجب رد فرض صفر می‌شود.



p-value	Evidence against H_0
$p > 0.10$	Weak or no evidence
$0.05 < p \leq 0.10$	Moderate evidence
$0.01 < p \leq 0.05$	Strong evidence
$p \leq 0.01$	Very strong evidence

محاسبه P مقدار

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$ p-Value=$P_{\theta_0}(X \geq x)$

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$ p-Value=$P_{\theta_0}(X \leq x)$

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$ p-Value=$2\min(P_{\theta_0}(X \leq x), P_{\theta_0}(X \geq x))$

در یک بازی شانسی، باید یک سکه پرتاب شود. اگر سکه شیر بیاید برنده خواهیم بود و در غیر اینصورت بازنده. برگزار کننده این بازی ادعا دارد که سکه‌اش ناریب است. یعنی احتمال ظاهر شدن شیر با خط برابر است. برای اینکه ادعای برگزار کننده را بررسی کنیم یک آزمون آماری تشکیل می‌دهیم.

اگر p احتمال مشاهده شیر باشد، فرضیه‌های این آزمون آماری به صورت زیر است:

$$\begin{cases} H_0 : p = \frac{1}{2} \\ H_1 : p > \frac{1}{2} \end{cases}$$

حال اگر Xرا تعداد شیر در ۱۰ بار پرتاب سکه در نظر بگیریم، با انجام این آزمایش، نتیجه آماره آزمون (یعنی همان X براساس نمونه تصادفی (شمارش تعداد شیرها در ۱۰ بار پرتاب سکه) برابر با ۶ شده است.

$$P(X > 6 |\ H_0) = 1 - P(X \leq 5 |\ p = \tfrac{1}{2}) =$$

$$1 - (\sum_{i=1}^{5} \binom{10}{i} \tfrac{1}{2}^{i} \times \tfrac{1}{2}^{10-i}) = 1 - 0.6230 = 0.3770$$

متغیر تصادفی تعداد زدگی‌ها در یک توپ پارچه، دارای توزیع پواسن با پارامتر λ است. طبق نظر کارشناس کارخانه متوسط تعداد زدگی در هر توپ پارچه برابر با ۵ است. به طور تصادفی یک توپ از پارچه‌ها انتخاب شده و تعداد زدگی‌ها برابر با ۱۰ شمارش شده است. در سطح خطای ۵٪، گفته کارشناس را بررسی می‌کنیم.

$$\begin{cases} H_0 : \lambda = 5 \\ H_1 : \lambda \neq 5 \end{cases}$$

$$X \sim P(\lambda)$$

$$P_{\lambda=5}(X \leq 10) = \sum^{10} \frac{e^{-5}5^k}{k!} = 0.9863$$

$$P_{\lambda=5}(X \geq 10) = 1 - P_{\lambda=5}(X \leq 9) = 1 - (\sum_{k=1}^{9} \frac{e^{-5}5^k}{k!}) = 1 - 0.9682 = 0.0318$$

$$\text{p-Value} = 2\min(P(X \leq 10), P(X \geq 10)) = 2\min(0.9863, 0.0318) = 0.0636$$

انواع آزمون‌ها آماری

آزمون پارامتری

۱)مشاهدات دارای توزیع مشخص باشند (در بیشتر آزمون‌های پارامتری باید توزیع جامعه‌ای که از آن نمونه‌گیری شده است، نرمال باشد)

۲)مشاهدات مستقل از هم باشند

* اگر شرایط آزمون پارامتری برقرار باشد، توان آزمون پارامتری بیشتر از آزمون ناپارامتری است.

آزمون ناپارامتری

آزمون نرمال بودن

- آزمون شاپیرو ویلک

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

* Vماتریس کوواریانس آماره‌های ترتیبی است.
* بردار m نیز مقدار امید ریاضی آماره‌های ترتیبی است که با میانگین تخمین زده شده‌است.

$$(a_1, \ldots, a_n) = \frac{m^{\mathsf{T}} V^{-1}}{C}$$

$$C = \|V^{-1}m\| = (m^{\mathsf{T}} V^{-1} V^{-1} m)^{1/2}$$

$$\overline{x} = (x_1 + \cdots + x_n)/n$$

آزمون‌های پارامتری

- آزمون تک نمونه‌ای میانگین جامعه
- برابری واریانس دو نمونه
- برابری میانگین دو جامعه برای نمونه‌های مستقل
- برابری میانگین دوجامعه زوجی

اداره بهداشت یک شهر، می‌خواهد تعیین کند که آیا میانگین تعداد باکتری‌ها، در واحد حجم آب شهر از سطح ایمنی یعنی ۲۰۰ بیشتر است یا نه. پژوهشگران ده نمونه از آب، هر یک به حجم واحد، را گرداوری کرده و تعداد باکتری موجود در هر نمونه را اندازه گرفتند. آیا داده‌های زیر موجب نگرانی اداره بهداشت می‌شود؟

Data = c(175,190,215,198,184,207,210,193,196,180)

داده‌ی میزان تولید دو نوع ذرت جمع‌آوری شده است. آیا میانگین‌های تولید این دو نوع ذرت تفاوت معناداری با هم دارند؟

داده‌های sleep میزان اثر دو نوع قرص خواب‌آور را بر روی ۲۰ نفر نشان می‌دهد. آیا میزان تاثیر دو نوع قرص با هم برابر است؟

آزمون‌های ناپارامتری

- ویلکاکسون با نمونه‌های مستقل
- یک نمونه‌ای ویلکاکسون
- آزمون ویلکاکسون با نمونه‌های زوجی

میزان حقوق دریافتی ده نفر از کارگران یک کارخانه در آمریکا جمع‌آوری شده است. یک فاصله اطمینان ۹۵ درصدی برای میزان حقوق دریافتی این کارگران به دست آورید.

Data = c(110, 12, 25, 98, 1017, 540, 54, 43,150, 432)

داده مربوط به میزان انرژی مصرف شده در گروه از زنان چاق و لاغر جمع‌آوری شده است. آیا می‌توان گفت میزان انرژی دو گروه با هم برابر است؟

میزان فشار خون ۱۵ مرد به وسیله دستگاه و کارشناس اندازه‌گیری شده است. آیا می‌توان گفت بین این دو روش اندازه‌گیری تفاوت معنی دار وجود دارد؟