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

An instructor has 50 exams that will be graded in sequence. The times required to grade the 50 exams are independent, with a common distribution that has mean 20 minutes and standard deviation 4 minutes. Approximate the probability that the instructor will grade at least 25 of the exams in the first 450 minutes of work.

$S_{25} = X_1 + \cdots + X_{25}$ $\quad X_i$: $E[X_i]=20$, $V[X_i]=16$

$E[S_{25}] = 500$ $\quad V[S_{25}] = 400$ $\quad S_{25} \sim N(500, 400)$

$$P[S_n \le 450] = P\left[\frac{S_{25}-500}{\sqrt{400}} \le \frac{450-500}{\sqrt{400}}\right] = P[Z \le -2.5]$$

$$= 1 - P[Z \le 2.5] = 1 - \Phi(2.5) = 0.006$$

یادآوری: $E(X) = \frac{N+1}{2}$ $V(X) = \frac{N^2-1}{12}$ — uniform توزیع متغیر گسسته

2. If 10 fair dice are rolled, find the approximate probability that the sum obtained is between 30 and 40, inclusive.

Sol: $X_i = \begin{cases} 1 & 1/6 \\ 2 & 1/6 \\ \vdots & \\ 6 & 1/6 \end{cases}$ $\quad E[X_i] = \frac{7}{2}$ $\quad V[X_i] = E[X_i^2] - (E[X_1])^2 = \frac{35}{12}$

Since $S_{10} = X_1 + \cdots + X_{10}$ is discrete, then one can compute the following $P[29.5 \le S_{10} \le 40.5]$ → چون گسسته است inclusive باشد

using $S_{10} \sim N\left(\frac{70}{2}, \frac{350}{12}\right)$. Thus

$$P[29.5 \le S_{10} \le 40.5] = P\left[\frac{29.5-35}{\sqrt{\frac{350}{12}}} \le Z \le \frac{40.5-35}{\sqrt{\frac{350}{12}}}\right] = P[-1.02 \le Z \le 1.02]$$

$$= \Phi(1.02) - (1 - \Phi(1.02)) = 0.85 - (1 - 0.85) = 0.70$$

3. The strong law of large numbers states that, with probability 1, the successive arithmetic averages of a sequence of independent and identically distributed random variables converge to their common mean μ. What do the successive geometric averages converge to? That is, what is

$$\lim_{n\to\infty} \left(\prod_{i=1}^{n} X_i\right)^{1/n}$$

Cousider $\left(\prod_{i=1}^{n} X_i\right)^{1/n}$, Take logarithm ...

$$\log\left[\left(\prod_{i=1}^{n} X_i\right)^{1/n}\right] = \frac{1}{n}\sum_{i=1}^{n} \log(X_i) = \frac{Y_1 + \cdots + Y_n}{n} = \bar{Y}_n$$

where $Y_i = \log(X_i)$

By the LLN, $\bar{Y}_n \to E[Y_i] = E[\log(X_i)]$ Thus

$$\log\left[\left(\prod_{i=1}^{n} X_i\right)^{1/n}\right] \xrightarrow{p} E[\log(X_i)]$$

So that

$$\left(\prod_{i=1}^{n} X_i\right)^{1/n} \xrightarrow{p} e^{E[\log(X_i)]}$$

4. A tobacco company claims that the amount of nicotine in one of its cigarettes is a random variable with mean 2.2 mg and standard deviation .3 mg. However, the average nicotine content of 100 randomly chosen cigarettes was 3.1 mg. Approximate the probability that the average would have been as high as or higher than 3.1 if the company's claims were true.

If the claim were true, then, by CLT,

$$\overline{X}_{100} = \frac{X_1 + \cdots + X_{100}}{100} \sim N(2.2, 0.0009)$$

$E[X_i] = 2.2 \qquad V[X_i] = 0.3^2 = 0.09 \qquad \overline{X}_{100} \sim N\left(E[X_i], \frac{V[X_i]}{100}\right)$

$$P[\overline{X}_{100} > 3.1] = P\left[Z > \frac{3.1 - 2.2}{\sqrt{0.0009}}\right] = P\left[Z > \frac{0.9}{0.03}\right] = P[Z > 5.196] \approx 0$$

$$\overline{X}_n = \frac{S_n}{n} \rightarrow X_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad n \geq 30$$

$$S_n = X_1 + X_2 + \cdots + X_n \rightarrow E(S_n) = n \cdot \mu$$

$$V(S_n) = n\sigma^2$$

5. Suppose that X is a random variable with mean and variance both equal to 20. What can be said about $P\{0 < X < 40\}$?

① $\mu = 20 \quad \sigma^2 = 20$. Let us apply the Chebyshev Inequality to $Y = X - \mu$.

$$E[Y] = 0$$
$$Var[Y] = Var[X-\mu] = Var[X] = \sigma^2 = 20$$
$$E[|Y|^2] = E[Y^2] = Var[Y] + E[Y]^2 = \sigma^2 + 0 = 20$$

$$P[0 < X < 40] = P[|X-\mu| < 20]$$
$$= P[|Y| < 20]$$
$$\geq 1 - \frac{E[|Y|^2]}{20^2} = 1 - \frac{20}{20^2} = 19/20$$

6. A person has 100 light bulbs whose lifetimes are independent exponentials with mean 5 hours. If the bulbs are used one at a time, with a failed bulb being replaced immediately by a new one, approximate the probability that there is still a working bulb after 525 hours.

7. An insurance company has 10,000 automobile policyholders. The expected yearly claim per policyholder is \$240, with a standard deviation of \$800. Approximate the probability that the total yearly claim exceeds \$2.7 million.

Note: (Corollary of Chebyshev Inequality)

$$\forall X: \quad P[|X-\mu| < a] \geq 1 - \frac{Var(X)}{a^2}$$

Markov Inequality: The same as Chebyshev for generic k

8. Let $X_1, \ldots, X_{20}$ be independent Poisson random variables with mean 1.

(a) Use the Markov inequality to obtain a bound on

$$P\left\{\sum_1^{20} X_i > 15\right\}$$

(b) Use the central limit theorem to approximate

$$P\left\{\sum_1^{20} X_i > 15\right\}.$$

② X_i = lifetime bulb i $\quad E[X_i] = 5 \quad V[X_i] = 25$

$S_{100} = X_1 + \cdots + X_{100}$

$S_{100} \sim N(100 \times 5, 100 \times 25) = N(500, 2500)$

$$P[S_{100} > 525] = P\left[Z > \frac{525-500}{\sqrt{2500}}\right] = 1 - \Phi(1/2)$$

$$= 1 - P(Z \leq 0.5) = 1 - \Phi(0.5)$$

9. Let X be a discrete random variable whose possible values are $1, 2, \ldots$. If $P\{X = k\}$ is nonincreasing in $k = 1, 2, \ldots$, prove that

$$P\{X = k\} \leq 2\frac{E[X]}{k^2}$$

QUIZ - CONVERGENCE AND LIMIT THEOREMS

Quiz 1 Given a sequence $\{X_n, n \in \mathbb{N}\}$ of random variables, we say that it converges in probability to a variable X iff

[a] $\lim_{n\to\infty} P[X_n \in A] = P[X \in A]$ for all $A \subseteq \mathbb{R}$.

[b] $\{\lim_{n\to\infty} P[X_n \in A] - P[X \in A]\} < \epsilon$ for all $\epsilon > 0$.

[c] $\lim_{n\to\infty} P[\{\omega : |X_n(\omega) - X(\omega)| < \epsilon\}] = 1$ for all $\epsilon > 0$.

[d] $P[\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}] = 1$ for all $\epsilon > 0$.

Quiz 2 Let $\{X_n, n \in \mathbb{N}\}$ be a sequence of non-negative random variables that converges in distribution to a non-negative variable X. Then, which of the following is also surely verified?

[a] $\lim_{n\to\infty} P[\{\omega : |X_n(\omega) - X(\omega)| < \epsilon\}] = 1$ for all $\epsilon > 0$.

[b] $\lim_{n\to\infty} E[X_n] = E[X]$.

[c] $\lim_{n\to\infty} E[\exp(-X_n)] = E[\exp(-X)]$.

[d] The sequence $\{Y_n, n \in \mathbb{N}\}$, where $Y_n = (X_n - X)$, converges in distribution to the constant 0.

Quiz 3 Let $\{X_n, n \in \mathbb{N}\}$ be a sequence of random variables that converges in distribution to a variable X. Then, which of the following is also surely verified?

[a] The sequence $\{X_n, n \in \mathbb{N}\}$ also converges in probability to X.

[b] The sequence $\{Y_n, n \in \mathbb{N}\}$, where $Y_n = |X_n - X|$, converges in distribution to $E[X]$.

[c] The sequence $\{\overline{X}_n, n \in \mathbb{N}\}$ converges to a $Z \sim N(0,1)$.

[d] None of the others.

Quiz 4 Let $Z \sim N(0,1)$. Then, by using the Chebishev's inequality one can affirm that:

[a] $P[Z^2 \geq a] \leq 1/a$ for all $a > 0$.

[b] $P[Z \geq a] \leq 1/a$ for all $a > 0$.

[c] $P[Z \geq a] \leq E[Z^2]/a^2$ for all $a > 0$.

[d] $P[Z^2 \geq a] \leq 1/a^2$ for all $a > 0$.

Quiz 5 Let $\{X_n, n \in \mathbb{N}\}$ be a sequence of independent random variables such that, for any n, $X_n \sim U[-1/n, 1/n]$. Let F_n denotes the distribution of X_n. Which all the following is true?

[a] $\lim_{n\to\infty} F_n(t) = t$ for all $t > 0$.

[b] $\lim_{n\to\infty} F_n(t) = 1$ for all $t > 0$.

[c] $\lim_{n\to\infty} F_n(t) = t$ for all $t \in \mathbb{R}$.

[d] $\lim_{n\to\infty} F_n(t) = 1$ for all $t \in \mathbb{R}$.

ANSWERS: Q1: c, Q2: c, Q3: d, Q4: a, Q5: b

Suppose that $Y_1 = 0.42$, $Y_2 = 0.10$, $Y_3 = 0.65$, and $Y_4 = 0.23$ is a random sample of size 4 from the pdf

$$f_Y(y;\theta) = \theta y^{\theta-1}, \quad 0 \leq y \leq 1$$

Find the method of moments estimate for θ.

The first theoretical moment of Y is $\frac{\theta}{\theta+1}$:

$$E(Y) = \int_0^1 y \cdot \theta y^{\theta-1}\, dy$$

$$= \theta \cdot \frac{y^{\theta+1}}{\theta+1}\Big|_0^1$$

$$= \frac{\theta}{\theta+1}$$

Setting $E(Y)$ equal to $\frac{1}{n}\sum_{i=1}^{n} y_i (= \overline{y})$, the first sample moment, gives

$$\frac{\theta}{\theta+1} = \overline{y}$$

which implies that the method of moments estimate for θ is

$$\theta_e = \frac{\overline{y}}{1-\overline{y}}$$

Here, $\overline{y} = \frac{1}{4}(0.42 + 0.10 + 0.65 + 0.23) = 0.35$, so

$$\theta_e = \frac{0.35}{1-0.35} = 0.54$$

---

suppose that $X_1 = 3$, $X_2 = 5$, $X_3 = 4$, and $X_4 = 2$ is a set of four independent observations representing the Poisson probability model,

$$p_X(k;\ \lambda) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

Find the maximum likelihood for λ.

$$L(\lambda) = e^{-\lambda}\frac{\lambda^3}{3!} \cdot e^{-\lambda}\frac{\lambda^5}{5!} \cdot e^{-\lambda}\frac{\lambda^4}{4!} \cdot e^{-\lambda}\frac{\lambda^2}{2!} = e^{-4\lambda}\lambda^{14}\frac{1}{3!5!4!2!}$$

Then $\ln L(\lambda) = -4\lambda + 14\ln\lambda - \ln(3!5!4!2!)$. Differentiating $\ln L(\lambda)$ with respect to λ gives

$$\frac{d\ln L(\lambda)}{d\lambda} = -4 + \frac{14}{\lambda}$$

To find the λ that maximizes $L(\lambda)$, we set the derivative equal to zero. Here $-4 + \frac{14}{\lambda} = 0$ implies that $4\lambda = 14$, and the solution to this equation is $\lambda = \frac{14}{4} = 3.5$.

⑤ Recall that $\sum_{i=1}^{K} i = \frac{K(K+1)}{2} \geq \frac{K^2}{2}$. Thus

$$E[X] = \sum_{i=1}^{\infty} i\, P[X=i] \geq \sum_{i=1}^{K} i\, P[X=i]$$

$$\geq \sum_{i=1}^{K} i \cdot P[X=K] \quad (\text{since } P[X=i] \downarrow \text{ in } i)$$

$$\geq P[X=K] \cdot \frac{K(K+1)}{2} \geq P[X=K] \cdot \frac{K^2}{2}$$

8\. Let $X_1, \ldots, X_{20}$ be independent Poisson random variables with mean 1.

(a) Use the Markov inequality to obtain a bound on

$$P\left\{\sum_1^{20} X_i > 15\right\}$$

(b) Use the central limit theorem to approximate

$$P\left\{\sum_1^{20} X_i > 15\right\}.$$

(a) $E[X_i] = Var(X_i) = 1$ $\qquad S_{20} = \sum^{20} X_i$ $\qquad E[S_{20}] = 20$

$V[S_{20}] = 20$

[illegible]

$E(X^2) = V(X) + [E(X)]^2 \leftrightarrow E[S_{20}^2] = 20 + 20^2 = 420$

$$P[S_{20} \geq 15] = P[S_{20} \geq 14]$$

$$\leq \frac{E[S_{20}^2]}{14^2} = \frac{420}{196} \text{ --- useless}$$

Standard Normal Probabilities

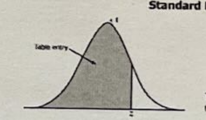


Table entry for z is the area under the standard normal curve to the left of z.

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

: exponensial برای MLE

In general:

$$L(\lambda)=\prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{k_i}}{k_i!}=e^{-n\lambda}\lambda^{\sum_{i=1}^{n}k_i}\frac{1}{\prod_{i=1}^{n}k_i!}$$

$$\ln L(\lambda)=-n\lambda+\left(\sum_{i=1}^{n}k_i\right)\ln\lambda-\ln\prod_{i=1}^{n}k_i!$$

$$\frac{d\ln L(\lambda)}{d\lambda}=-n+\frac{\sum_{i=1}^{n}k_i}{\lambda}$$

Setting the derivative equal to 0 gives

$$-n+\frac{\sum_{i=1}^{n}k_i}{\lambda}=0 \quad \text{which implies that } \lambda_e=\frac{\sum_{i=1}^{n}k_i}{n}=\bar{k}.$$

Setting the derivative of $\ln L(\theta)$ equal to 0 gives

$$\frac{d\ln L(\theta)}{d\theta}=\frac{-2n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}y_i=0$$

which implies that

$$\theta_e=\frac{1}{2n}\sum_{i=1}^{n}y_i$$

$$\sum_{i=1}^{5}y_i=9.2+5.6+18.4+12.1+10.7=56.0,$$

$$\theta_e=\frac{1}{2(5)}(56.0)=5.6$$

12. Suppose an isolated weather-reporting station has an electronic device whose time to failure is given by the exponential model

$$f_Y(y;\ \theta)=\frac{1}{\theta}e^{-y/\theta},\quad 0\le y<\infty;\ 0<\theta<\infty$$

The station also has a spare device, so the time until this instrument is not available is the sum of these two exponential pdfs, which is

$$f_Y(y;\ \theta)=\frac{1}{\theta^2}ye^{-y/\theta},\quad 0\le y<\infty;\ 0<\theta<\infty$$

Five data points have been collected—9.2, 5.6, 18.4, 12.1, and 10.7. Find the maximum likelihood estimate for θ.

$$L(\theta)=\prod_{i=1}^{n}\frac{1}{\theta^2}y_ie^{-y_i/\theta}=\theta^{-2n}\left(\prod_{i=1}^{n}y_i\right)e^{-(1/\theta)\sum_{i=1}^{n}y_i}$$

13.

1 Exercise 1: Poisson data

Let $X_1,\cdots,X_n$ a sample, recording the number of yearly claims (for n consecutive years) that a condominium have issued to an insurance company, for some adverse event that could be covered by its policy. Suppose that this data-set can be modeled as a simple random sample from the Poisson distribution. (a) Find a method of moment estimator (b) Find the MLE. In order to perform Bayesian inference, adopt a Gamma prior and prove that this family of distributions is conjugate to the Poisson family. Suppose that the company experience is that most of the policyholders issue one or two claims per year (rarely 0 or 3, other cases have negligible probabilities). Try to encode that information into the choice of the prior, and exhibit a Bayesian point estimator that takes this information into account. What can you say about the distribution of this estimators?

poisson: $E[X]=\lambda$ $\quad M_1 = E[X] = \bar{X}_n = \lambda(x)$ → بعد poisson

MME) $\lambda = E[X] \longrightarrow \lambda = \bar{X}_n \quad (\bar{X}_n = M_1)$ ⑤

MLE) Let $\bar{X}=(X_1,\dots,\bar{X}_n)$. Then

$$L(\lambda|\bar{x}) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum x_i} \cdot e^{-n\lambda}}{\prod_i x_i!}$$

$$\ln L(\lambda|\bar{x}) = -n\lambda + \sum_{i=1}^{n} x_i \ln(\lambda) - \ln\left(\prod x_i!\right)$$

$$0 = \frac{d \ln L(\lambda|\bar{x})}{d\theta} = -n + \frac{\sum_i x_i}{\lambda} \longrightarrow \hat{\lambda} = \frac{\sum x_i}{n} = \bar{X}_n$$

BAYES)

Assume the prior to be $\sim \Gamma(\alpha,\gamma)$, i.e.

$$\pi(\lambda) = \frac{\gamma e^{-\gamma\lambda}(\gamma\lambda)^{\alpha-1}}{\Gamma(\alpha)} = \frac{\gamma^{\alpha}\lambda^{\alpha-1}e^{-\gamma\lambda}}{\Gamma(\alpha)}$$

Let $S = X_1 + \dots + X_n$. $S \sim \text{Poiss}(n\lambda) \rightarrow f_S(s|\lambda) = \frac{(n\lambda)^s}{s!} e^{-n\lambda}$

Thus

$$\pi(\lambda|s) \sim f_S(s|\lambda)\cdot\pi(\lambda)$$

$$= \frac{(n\lambda)^s}{s!} e^{-n\lambda} \cdot \frac{\gamma^{\alpha}\lambda^{\alpha-1}e^{-\gamma\lambda}}{\Gamma(\alpha)}$$

$$= \underbrace{\frac{n^s}{s!}\frac{\gamma^{\alpha}}{\Gamma(\alpha)}}_{\text{constant in } \lambda} \cdot \underbrace{e^{-(n+\gamma)\lambda}\cdot\lambda^{s+\alpha-1}}_{\text{terms in } \lambda \rightsquigarrow \Gamma(s+\alpha,\, n+\gamma)}$$

The Bayes estimate of λ is $E[\Gamma(s+\alpha, n+\gamma)]$, i.e.,

$$\hat{\lambda} = \frac{s+\alpha}{n+\gamma} \quad (s=\textstyle\sum x_i) \qquad \left(\text{remember } E[\Gamma(\alpha,\gamma)] = \alpha/\gamma\right)$$

By experience $E[\pi] \cong 1.5$, $V[\pi] \cong 1$, since $E[\Gamma(\alpha,\gamma)] = \alpha/\gamma$ and $V[\Gamma(\alpha,\gamma)] = \alpha/\gamma^2$, then

$$\begin{cases} \alpha/\gamma = \frac{3}{2} \\ \alpha/\gamma^2 = 1 \end{cases} \rightarrow \alpha = 9/4 \quad \gamma = 3/2$$

H.

2 Exercise 2: Exponential data

Let $T_1, \cdots, T_n$ be a sample, recording the uptimes of a server. An uptime is a measure of system reliability, expressed as the consecutive time a machine, typically a computer, has been working and available. Let us suppose our sample is a simple random sample from the Exponential distribution. Find a method of moment estimator. Find the MLE. In order to perform Bayesian inference, adopt a Gamma prior, and prove that this family of distributions is conjugate to the Exponential family.

For you ---

MME) $M_1 = E(x) = \frac{1}{\lambda} = \bar{X}_n \Rightarrow \lambda = \frac{1}{\bar{X}_n}$

MLE) $L(x|\lambda) = \prod \lambda e^{-\lambda x} = \lambda^n e^{-n\lambda\sum x_i} \xrightarrow{\ln} n\ln\lambda - n\lambda\sum x_i$

$\xrightarrow{d/d\lambda} \frac{n}{\lambda} - n\sum x_i = 0 \rightarrow \lambda = \frac{1}{\sum x_i}$?

5, ?

15. Let $X_1, \ldots, X_n$ be iid with pdf

$$f(x|\theta) = \theta x^{\theta-1}, \quad 0 \le x \le 1, \quad 0 < \theta < \infty.$$

Find the MLE of θ.

$$L(\theta|\vec{x}) = \theta^n \prod_{i=1}^{n} x_i^{\theta-1}$$

$$\ln L(\theta|\vec{x}) = n \ln\theta + (\theta-1)\ln\left(\prod x_i\right) = n\ln\theta + (\theta-1)\sum \ln(x_i)$$

$$\frac{d \ln(L(\theta|\vec{x}))}{d\theta} = \frac{n}{\theta} + \sum \ln(x_i) = 0 \qquad \hat{\theta}_n = -\frac{n}{\sum^n \ln(x_i)}$$

16. Each item produced will, independently, be defective with probability p. If the prior distribution on p is uniform on (0, 1), compute the posterior probability that p is less than .2 given

(a) a total of 2 defectives out of a sample of size 10;
(b) a total of 1 defective out of a sample of size 10;
(c) a total of 10 defectives out of a sample of size 10.

$X \sim \text{Bern}(p)$. $\quad \pi(p) = \begin{cases} 1 & p \in (0,1) \\ 0 & p \notin (0,1) \end{cases}$

$$\pi(p|\vec{x}) = L(p|\vec{x}) \cdot \pi(p)$$

$$\propto p^{\sum x_i} (1-p)^{n-\sum x_i} \, \mathbb{1}_{(0,1)}(p)$$

(a) $$\frac{\int_0^{0.2} p^2(1-p)^8 dp}{\int_0^1 p^2(1-p)^8 dp} = \frac{\left[-\frac{(1-p)^9}{9} + \frac{(1-p)^{10}}{5} - \frac{(1-p)^{11}}{11}\right]_0^{0.2}}{\left[-\frac{(1-p)^9}{9} + \frac{(1-p)^{10}}{5} - \frac{(1-p)^{11}}{11}\right]_0^{1}} = 0.61$$

$$E[x] = \bar{Y} = \int_0^1 y \cdot (\theta^2+\theta) y^{\theta-1}(1-y) = \theta^2 + \theta \int_0^1 y^\theta - y^{\theta+1} dy = \theta(\theta+1)\left(\frac{\theta+2-\theta-1}{(\theta+1)(\theta+2)}\right)$$

$y\theta$

$$\frac{y^{\theta+1}}{\theta+1} - \frac{y^{\theta+2}}{\theta+2}$$

17. **5.2.17.** Use the method of moments to estimate θ in the pdf

$$f_Y(y;\theta) = (\theta^2+\theta)y^{\theta-1}(1-y), \quad 0 \le y \le 1$$

Assume that a random sample of size n has been collected.

Ans: $\Rightarrow \frac{\theta}{\theta+2} = \bar{Y}$

$$\frac{2\bar{Y}}{1-\bar{Y}}$$

$$\theta = \theta\bar{Y} + 2\bar{Y} \rightarrow \theta = \frac{2\bar{Y}}{1-\bar{Y}}$$

18. **5.2.20.** Suppose that $Y_1 = 8.3$, $Y_2 = 4.9$, $Y_3 = 2.6$, and $Y_4 = 6.5$ is a random sample of size 4 from the two-parameter uniform pdf,

$$f_Y(y;\theta_1,\theta_2) = \frac{1}{2\theta_2}, \quad \theta_1 - \theta_2 \le y \le \theta_1 + \theta_2$$

Use the method of moments to calculate $\hat{\theta}_{1e}$ and $\hat{\theta}_{2e}$.

easy

$Y \sim U[\theta_1-\theta_2, \theta_1+\theta_2]$ uniform: $E(x) = \frac{1}{2}(a+b)$

$$E[Y] = \frac{(\theta_1-\theta_2)+(\theta_1+\theta_2)}{2} = \theta_1$$

$$V[Y] = \frac{[(\theta_1+\theta_2)-(\theta_1-\theta_2)]^2}{12} = \frac{(2\theta_2)^2}{12} = \frac{\theta_2^2}{3}$$

$$E[Y^2] = V[Y] + (E[Y])^2 = \frac{\theta_2^2}{3} + \theta_1^2 = \frac{1}{3}\left(\theta_2^2 + 3\theta_1^2\right)$$

Solve $\begin{cases} M_1 = \hat{\theta}_1 \\ M_2 = \frac{1}{3}\left(\hat{\theta}_2^2 + 3\hat{\theta}_1^2\right) \end{cases}$

where $M_1 = \frac{8.3+4.9+2.6+6.5}{4} = 5.575$

$$M_2 = \frac{(8.3)^2+(4.9)^2+(2.6)^2+(6.5)^2}{4} = 35.4775$$

$$\Rightarrow \hat{\theta}_1 = 5.575 \qquad \hat{\theta}_2 = \sqrt{3M_2 - 3\hat{\theta}_1^2} \approx 3.63$$

poisson : Fru - M

5.2.21. Find a formula for the method of moments estimate for the parameter θ in the Pareto pdf,

$$f_Y(y;\theta)=\theta k^\theta \left(\frac{1}{y}\right)^{\theta+1}, \quad y\ge k; \quad \theta\ge 1$$

Assume that k is known and that the data consist of a random sample of size n.

Ans: $\bar{y}/(\bar{y}-k)$

5.2.22. Calculate the method of moments estimate for the parameter θ in the probability function

$$p_X(k;\theta)=\theta^k(1-\theta)^{1-k}, \quad k=0,1$$

if a sample of size 5 is the set of numbers 0, 0, 1, 0, 1. $\rightsquigarrow \bar{X}=2/5$

$E(x)=\sum xp_x$

$P_x(k=0)=1-\theta$

$P_x(k=1)=\theta$

$\rightarrow E(x)=\sum_{k=0,1} kp(x)=0(1-\theta)+\theta=\theta=\bar{X} \rightarrow \theta=0.4$

5.2.1 $\rightarrow \sum_{k=0,1} kP(k)=0+\theta=\theta=E(x)=\bar{X} \Rightarrow \theta=5/8$

5.2.3: $L(y|\lambda)=\pi\lambda e^{-\lambda y}=\lambda^n e^{-\lambda\sum y_i} \xrightarrow{\ln} n\ln\lambda-\lambda\sum y_i \xrightarrow{d/d\lambda} n/\lambda-\sum y_i=0$

$\Rightarrow \lambda=n/\sum y_i=4/32.8=0.121$ ✓

5.2.1. A random sample of size 8—$X_1=1, X_2=0, X_3=1, X_4=1, X_5=0, X_6=1, X_7=1$, and $X_8=0$—is taken from the probability function

$$p_X(k;\theta)=\theta^k(1-\theta)^{1-k}, \quad k=0,1; \quad 0<\theta<1$$

Find the maximum likelihood estimate for θ.

Ans: 5/8

5.2.3. Use the sample $Y_1=8.2, Y_2=9.1, Y_3=10.6$, and $Y_4=4.9$ to calculate the maximum likelihood estimate for λ in the exponential pdf

$$f_Y(y;\lambda)=\lambda e^{-\lambda y}, \quad y\ge 0$$

Ans: 0.122

5.2.4. Suppose a random sample of size n is drawn from the probability model

$$p_X(k;\theta)=\frac{\theta^{2k}e^{-\theta^2}}{k!}, \quad k=0,1,2,\ldots$$

Find a formula for the maximum likelihood estimator, $\hat{\theta}$.

Ans $\sqrt{\frac{\sum k_i}{n}}$

5.2.4: $L(k|\theta)=\pi\frac{\theta^{2k}e^{-\theta^2}}{k!}=\theta^{2\sum k_i}e^{-n\theta^2}/\pi k! \xrightarrow{\ln} 2\sum k_i\ln\theta-n\theta^2+\ln(1/\pi k!)$

$\xrightarrow{d/d\lambda} \frac{2\sum k_i}{n}-2n\theta=0 \rightsquigarrow \theta=\sqrt{\sum k_i/n}$

5.2.5. Given that $Y_1=2.3, Y_2=1.9$, and $Y_3=4.6$ is a random sample from

$$f_Y(y;\theta)=\frac{y^3e^{-y/\theta}}{6\theta^4}, \quad y\ge 0$$

calculate the maximum likelihood estimate for θ.

Ans: 0.733

5.2.7. An engineer is creating a project scheduling program and recognizes that the tasks making up the project are not always completed on time. However, the completion proportion tends to be fairly high. To reflect this condition, he uses the pdf

$$f_Y(y;\theta)=\theta y^{\theta-1}, \quad 0\le y\le 1, \quad \text{and} \quad 0<\theta$$

where y is the proportion of the task completed. Suppose that in his previous project, the proportions of tasks completed were 0.77, 0.82, 0.92, 0.94, and 0.98. Estimate θ.

Ans: 8.00

5.2.11. Find the maximum likelihood estimate for θ in the pdf

$$f_Y(y;\theta)=\frac{2y}{1-\theta^2}, \quad \theta\le y\le 1$$

if a random sample of size 6 yielded the measurements 0.70, 0.63, 0.92, 0.86, 0.43, and 0.21.

Ans: $\min\{Y_i\}$ (rws: 0.21) (see below for details)

5.8.1: $f(\theta|x)\propto f(x|\theta)\cdot f(\theta)$ → the prior distribution of θ.

likelihood function

5.8.1. Suppose that X is a geometric random variable, where $p_X(k|\theta)=(1-\theta)^{k-1}\theta, k=1,2,\ldots$. Assume that the prior distribution for θ is the beta pdf with parameters r and s. Find the posterior distribution for θ.

$\rightarrow f(\theta)=\theta^{r-1}(1-\theta)^{s-1}/B(r,s)$

Ans: The numerator of $\pi_\theta(\theta|X=k)$ is

$p_X(k|\theta)\cdot f_\theta(\theta)=\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\underbrace{\theta^r(1-\theta)^{s+k-2}}_{\text{variable part of } \nu\beta(r+1,s+k-1)}$

5.8.17 $f(\theta|x) \propto f(x|\theta) f(\theta)$

$f(x|\theta) = \theta(1-\theta)^{x-1}$

$f(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \theta^{r-1} (1-\theta)^{s-1}$ → normalization constant

$\Rightarrow f(\theta|x) = \theta(1-\theta)^{x-1} \theta^{r-1} (1-\theta)^{s-1} + C$

$= \theta^{r} (1-\theta)^{x+s-2} + C$

این هم یک توزیع بتا با پارامترهای $r+1$ و $x+s-1$ است

$B(r+1, x+s-1)$

Let $X_1, \cdots, X_n$ be a sample from the geometric distribution with parameter p.

- Write down the likelihood function of the sample
- Calculate the maximum likelihood estimator
- What can you say about the asymptotic distribution of the estimator? Is the estimator consistent?

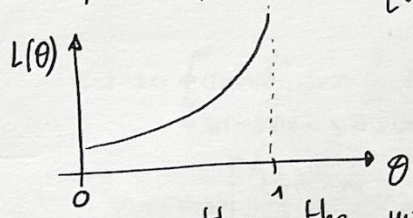
$f(k) = p(1-p)^{k-1}$

$L(k_i|p) = \prod p(1-p)^{k_i-1} = p^n (1-p)^{\sum k_i - n}$ $\xrightarrow{\ln}$ $n\ln p + (\sum k_i - n)\ln(1-p)$

$\frac{d}{dp} \rightarrow \frac{n}{p} - \frac{\sum k_i - n}{1-p} = 0 \rightarrow n - np = \sum k_i p - np \rightarrow p = \frac{n}{\sum k_i}$

You must consider it as a function of θ, with the y_i fixed constant.

Consider only the first part $\left[\frac{2^n}{(1-\theta^2)^n} \cdot \prod^n y_i\right]$



It is increasing, thus the max is in 1

But... could θ be greather than the y_i?

Consider $\prod_{i=1}^n \mathbb{1}_{[\theta,1]}(y_i)$

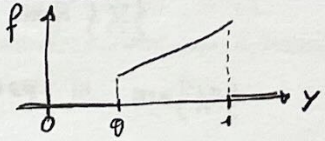
5.2.11. Find the maximum likelihood estimate for θ in the pdf

$$f_Y(y;\theta) = \frac{2y}{1-\theta^2}, \quad \theta \le y \le 1$$

if a random sample of size 6 yielded the measurements 0.70, 0.63, 0.92, 0.86, 0.43, and 0.21.

Ans: $\min\{y_i\}$ (Thus: 0.21)

$f(y,\theta) = \frac{2y}{1-\theta^2} \cdot \mathbb{1}_{[\theta,1]}(y)$



$L(\theta|\bar{x}) = \prod_{i=1}^n f(y_i,\theta)$

$= \frac{2^n}{(1-\theta^2)^n} \cdot \prod^n y_i \cdot \prod_{i=1}^n \mathbb{1}_{[\theta,1]}(y_i)$

If $\theta > y_i$ for at least one y_i (say, eg, if $\theta > y_1$) then $\mathbb{1}_{[\theta,1]}(y_1) = 0$, and the product becomes 0.

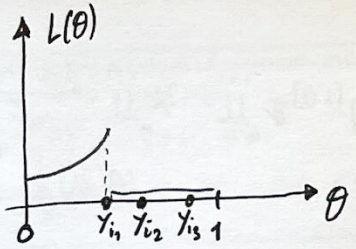
Thus θ must be smaller than all the y_i, i.e., it must be

$$\theta \le y_i \ \forall i$$

i.e.,

$$\theta \le \min(y_i)$$

Because of it, the correct graph of $L(\theta)$ is



and the max $L(\theta)$ is attained in $\min\{Y_i\}$.

$\hat{y}$ = MLE = $\arg\max L(\theta) = \min\{Y_i\}$

a. Suppose a coin, for which $p = P(\text{heads})$ is unknown, is to be tossed ten times for the purpose of estimating p with the function $\hat{p} = \frac{X}{10}$, where X is the observed number of heads. *If* $p = 0.60$, what is the probability that $\left|\frac{X}{10} - 0.60\right| \le 0.10$? That is, what are the chances that the estimator will fall within 0.10 of the true value of the parameter? Here $\hat{p}$ is discrete—the only values $\frac{X}{10}$ can take on are $\frac{0}{10}, \frac{1}{10}, \ldots, \frac{10}{10}$.

$$p_{\hat{p}}\left(\frac{k}{10}\right) = P\left(\hat{p} = \frac{k}{10}\right) = P(X = k) = \binom{10}{k}(0.60)^k(0.40)^{10-k}, \quad k = 0, 1, \ldots, 10$$

Therefore,

$$P\left(\left|\frac{X}{10} - 0.60\right| \le 0.10\right) = P\left(0.60 - 0.10 \le \frac{X}{10} \le 0.60 + 0.10\right)$$
$$= P(5 \le X \le 7)$$
$$= \sum_{k=5}^{7} \binom{10}{k}(0.60)^k(0.40)^{10-k}$$
$$= 0.6665$$

b. How likely is the estimator $\frac{X}{n}$ to lie within 0.10 of p if the coin in part (a) is tossed *one hundred* times?

$$P\left(\left|\frac{X}{100} - 0.60\right| \le 0.10\right) = P\left(0.50 \le \frac{X}{100} \le 0.70\right)$$
$$= P\left[\frac{0.50 - 0.60}{\sqrt{\frac{(0.60)(0.40)}{100}}} \le \frac{X/100 - 0.60}{\sqrt{\frac{(0.60)(0.40)}{100}}} \le \frac{0.70 - 0.60}{\sqrt{\frac{(0.60)(0.40)}{100}}}\right]$$
$$\doteq P(-2.04 \le Z \le 2.04)$$
$$= 0.9586$$

It was mentioned ~~at the outset of this section~~ that $\hat{\theta}_1 = \frac{3}{2}\overline{Y}$ and $\hat{\theta}_2 = Y_{\max}$ are two estimators for θ in the pdf $f_Y(y;\theta) = \frac{2y}{\theta^2}$, $0 \le y \le \theta$. Are either or both unbiased?

$$E(\hat{\theta}_1) = E\left(\frac{3}{2}\overline{Y}\right) = \frac{3}{2}E(\overline{Y}) = \frac{3}{2}E(Y) = \frac{3}{2}\cdot\frac{2}{3}\theta = \theta$$ ← unbiased

$$E[Y] = \int_0^\theta \frac{2y^2}{\theta^2}\,dy = \frac{2}{3}\,y^3/\theta^2\Big|_0^\theta = \frac{2\theta^3}{3\theta^2} = \frac{2}{3}\theta$$

$f_{Y_{\max}}(y) = n F_Y(y)^{n-1} f_Y(y)$

The cdf for Y is

$$F_Y(y) = \int_0^y \frac{2t}{\theta^2}\,dt = \frac{y^2}{\theta^2}$$

Then

$$f_{Y_{\max}}(y) = n\left(\frac{y^2}{\theta^2}\right)^{n-1}\frac{2y}{\theta^2} = \frac{2n}{\theta^{2n}}y^{2n-1},\ 0 \le y \le \theta$$

Therefore,

$$E(Y_{\max}) = \int_0^\theta y\cdot\frac{2n}{\theta^{2n}}y^{2n-1}\,dy = \frac{2n}{\theta^{2n}}\int_0^\theta y^{2n}\,dy = \frac{2n}{\theta^{2n}}\cdot\frac{\theta^{2n+1}}{2n+1} = \frac{2n}{2n+1}\theta$$

$\lim_{n\to\infty}\frac{2n}{2n+1}\theta = \theta$. Intuitively, this decrease in the bias makes sense because f_{θ_2} becomes increasingly concentrated around θ as n grows. ■

to estimator ... unbiased

Let $Y_1, \ldots, Y_n$ be a random sample from the pdf $f_Y(y;\theta) = \frac{2y}{\theta^2}$, $0 \le y \le \theta$. We know ~~from Example 5.4.2~~ that $\hat{\theta}_1 = \frac{3}{2}\overline{Y}$ and $\hat{\theta}_2 = \frac{2n+1}{2n}Y_{\max}$ are both unbiased for θ. Which estimator is more efficient?

$$E(Y^2) = \int_0^\theta y^2\cdot\frac{2y}{\theta^2}\,dy = \frac{2}{\theta^2}\int_0^\theta y^3\,dy = \frac{2}{\theta^2}\cdot\frac{\theta^4}{4} = \frac{1}{2}\theta^2$$

and

$$\mathrm{Var}(Y) = E(Y^2) - E(Y)^2 = \frac{1}{2}\theta^2 - \left(\frac{2}{3}\theta\right)^2 = \frac{\theta^2}{18}$$

Then

$$\mathrm{Var}(\hat{\theta}_1) = \mathrm{Var}\left(\frac{3}{2}\overline{Y}\right) = \frac{9}{4}\mathrm{Var}(\overline{Y}) = \frac{9}{4}\frac{\mathrm{Var}(Y)}{n} = \frac{9}{4n}\cdot\frac{\theta^2}{18} = \frac{\theta^2}{8n}$$

$\mathrm{Var}(\overline{Y}) = \frac{\mathrm{Var}(Y)}{n}$ از CLT

To address the variance of $\hat{\theta}_2 = \frac{2n+1}{2n}Y_{\max}$, we start with finding the variance of $Y_{\max}$. Recall that its pdf is

$$n F_Y(y)^{n-1} f_Y(y) = \frac{2n}{\theta^{2n}}y^{2n-1},\ 0 \le y \le \theta$$

From that expression, we obtain

$$E(Y_{\max}^2) = \int_0^\theta y^2\cdot\frac{2n}{\theta^{2n}}y^{2n-1}\,dy = \frac{2n}{\theta^{2n}}\int_0^\theta y^{2n+1}\,dy = \frac{2n}{\theta^{2n}}\cdot\frac{\theta^{2n+2}}{2n+2} = \frac{n}{n+1}\theta^2$$

and then

$$\mathrm{Var}(Y_{\max}) = E(Y_{\max}^2) - E(Y_{\max})^2 = \frac{n}{n+1}\theta^2 - \left(\frac{2n}{2n+1}\theta\right)^2 = \frac{n}{(n+1)(2n+1)^2}\theta^2$$

Finally,

$$\mathrm{Var}(\hat{\theta}_2) = \mathrm{Var}\left(\frac{2n+1}{2n}Y_{\max}\right) = \frac{(2n+1)^2}{4n^2}\mathrm{Var}(Y_{\max}) = \frac{(2n+1)^2}{4n^2}\cdot\frac{n}{(n+1)(2n+1)^2}\theta^2$$

$$= \frac{1}{4n(n+1)}\theta^2$$

Suppose the random variables $X_1, X_2, \ldots, X_n$ denote the number of successes (0 or 1) in each of n independent trials, where $p = P(\text{Success occurs at any given trial})$ is an unknown parameter. Then

$$p_{X_i}(k; p) = p^k (1-p)^{1-k}, \quad k = 0, 1; \quad 0 < p < 1$$

Let $X = X_1 + X_2 + \cdots + X_n$ = total number of successes and define $\hat{p} = \frac{X}{n}$. Clearly, $\hat{p}$ is unbiased for $p \left[E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{E(X)}{n} = \frac{np}{n} = p \right]$. How does $\text{Var}(\hat{p})$ compare with the Cramér-Rao lower bound for $p_{X_i}(k; p)$?

Note, first, that

$$\text{Var}(\hat{p}) = \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\text{Var}(X) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

(since X is a binomial random variable). To evaluate, say, the *second* form of the Cramér-Rao lower bound, we begin by writing

$$\ln p_{X_i}(X_i; p) = X_i \ln p + (1 - X_i) \ln(1-p)$$

Moreover,

$$\frac{\partial \ln p_{X_i}(X_i; p)}{\partial p} = \frac{X_i}{p} - \frac{1 - X_i}{1-p}$$

and

$$\frac{\partial^2 \ln p_{X_i}(X_i; p)}{\partial p^2} = -\frac{X_i}{p^2} - \frac{1 - X_i}{(1-p)^2}$$

Taking the expected value of the second derivative gives

$$E\left[\frac{\partial^2 \ln p_{X_i}(X_i; p)}{\partial p^2}\right] = -\frac{p}{p^2} - \frac{(1-p)}{(1-p)^2} = -\frac{1}{p(1-p)}$$

The Cramér-Rao lower bound, then, reduces to

$$\frac{1}{-n\left[-\frac{1}{p(1-p)}\right]} = \frac{p(1-p)}{n}$$

which *equals* the variance of $\hat{p} = \frac{X}{n}$. It follows that $\frac{X}{n}$ is the preferred statistic for estimating the binomial parameter p: No unbiased estimator can possibly be more precise.

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from the uniform pdf

$$f_Y(y; \theta) = \frac{1}{\theta}, \quad 0 \le y \le \theta$$

and let $\hat{\theta}_n = Y_{\max}$. We already know that $Y_{\max}$ is biased for θ, but is it consistent?

$$f_{Y_{\max}}(y) = \frac{n y^{n-1}}{\theta^n}, \quad 0 \le y \le \theta$$

Therefore,

$$P(|\hat{\theta}_n - \theta| < \varepsilon) = P(\theta - \varepsilon < \hat{\theta}_n < \theta) = \int_{\theta-\varepsilon}^{\theta} \frac{n y^{n-1}}{\theta^n}\, dy = \frac{y^n}{\theta^n}\Big|_{\theta-\varepsilon}^{\theta}$$
$$= 1 - \left(\frac{\theta - \varepsilon}{\theta}\right)^n$$

Since $[(\theta - \varepsilon)/\theta] < 1$, it follows that $[(\theta - \varepsilon)/\theta]^n \to 0$ as $n \to \infty$. Therefore, $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1$, proving that $\hat{\theta}_n = Y_{\max}$ is consistent for θ.

Given a population X, let θ be an unknown parameter of X (like the mean, or the variance). To estimate it, we can provide a sample $(X_1, X_2, ..., X_n)$, and define an estimator Θ_n.

a) The estimator Θ_n of θ is a function of the sample $(X_1, X_2, ..., X_n)$. [V] [F]
b) The estimator Θ_n of θ is a real number. [V] [F]
c) Θ_n is an estimator of θ if and only $E[\Theta_n] = \theta$. [V] [F]
d) The estimator Θ_n of is said to be consistent if $E[\Theta_n] = \theta$. [V] [F]
e) The estimator Θ_n of is said to be consistent if $E[\Theta_n] = 0$. [V] [F]
f) The estimator Θ_n of is said to be unbiased if $E[\Theta_n] = 0$. [V] [F]
g) The estimator Θ_n of is said to be unbiased if $E[\Theta_n] = \theta$. [V] [F]
h) The estimator Θ_n of is said to be consistent if $\lim_{n\to\infty} E[\Theta_n] = \theta$. [V] [F]
i) If θ is the mean, then the estimator $\Theta_n = \frac{X_1 + X_2 + ... X_n}{n}$ is unbiased. [V] [F]
l) If θ is the mean, then the estimator $\Theta_n = X_n$ is unbiased. [V] [F]
m) If θ is the mean, then the estimator $\Theta_n = \frac{X_1 + X_2 + ... X_n}{n}$ is consistent. [V] [F]
n) If θ is the mean, then the estimator $\Theta_n = X_n$ is consistent. [V] [F]
o) If θ is the variance, then the estimator $\Theta_n = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{n}$ is unbiased. [V] [F]
p) If θ is the variance, then the estimator $\Theta_n = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{n-1}$ is unbiased. [V] [F]

5.4.1) $\hat{\theta} = \frac{X_1 + X_2}{2}$ — sampling with replacement

$P(|\hat{\theta} - 3| > 1) = \frac{1}{10} + \frac{1}{10} = 0.2$

1,2 ⇒ 1.5 2,3 ⇒ 2.5 3,4 ⇒ 3.5 4,5 ⇒ 4.5

1,3 ⇒ 2 2,4 ⇒ 3 3,5 ⇒ 4

1,4 ⇒ 2.5 2,5 ⇒ 3.5

1,5 ⇒ 3

5.4.1. Two chips are drawn without replacement from an urn containing five chips, numbered 1 through 5. The average of the two drawn is to be used as an estimator, $\hat{\theta}$, for the true average of all the chips ($\theta = 3$). Calculate $P(|\hat{\theta} - 3| > 1.0)$.

Ans: 2/10

5.4.4. A sample of size $n = 16$ is drawn from a normal distribution where $\sigma = 10$ but μ is unknown. If $\mu = 20$, what is the probability that the estimator $\hat{\mu} = \overline{Y}$ will lie between 19.0 and 21.0? $P(19 < \hat{\mu} < 21)$

Ans: 0.31

5.4.9. A random sample of size 2, Y_1 and Y_2, is drawn from the pdf

$$f_Y(y;\theta) = 2y\theta^2, \quad 0 < y < \frac{1}{\theta}$$

What must c equal if the statistic $c(Y_1 + 2Y_2)$ is to be an unbiased estimator for $\frac{1}{\theta}$?

Ans: 1/2

5.4.11. Suppose that W is an unbiased estimator for θ. Can W^2 be an unbiased estimator for θ^2?

Ans: Only if W is constant

5.4.17. Let $X_1, X_2, \ldots, X_n$ denote the outcomes of a series of n independent trials, where

$$X_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

for $i = 1, 2, \ldots, n$. Let $X = X_1 + X_2 + \cdots + X_n$.

(a) Show that $\hat{p}_1 = X_1$ and $\hat{p}_2 = \frac{X}{n}$ are unbiased estimators for p.

(b) Intuitively, $\hat{p}_2$ is a better estimator than $\hat{p}_1$ because $\hat{p}_1$ fails to include any of the information about the parameter contained in trials 2 through n. Verify that speculation by comparing the variances of $\hat{p}_1$ and $\hat{p}_2$.

Ans:
(a) easy
(b)
$\mathrm{Var}(\hat{p}_1) = p(1-p)$
$\mathrm{Var}(\hat{p}_2) = \frac{p(1-p)}{n}$

5.4.19. Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from the pdf $f_Y(y;\theta) = \frac{1}{\theta}e^{-y/\theta}$, $y > 0$.

(a) Show that $\hat{\theta}_1 = Y_1, \hat{\theta}_2 = \overline{Y}$, and $\hat{\theta}_3 = n \cdot Y_{\min}$ are all unbiased estimators for θ.

(b) Find the variances of $\hat{\theta}_1, \hat{\theta}_2$, and $\hat{\theta}_3$.

(c) Calculate the relative efficiencies of $\hat{\theta}_1$ to $\hat{\theta}_3$ and $\hat{\theta}_2$ to $\hat{\theta}_3$.

Ans: (c)
$\frac{V(\hat{\theta}_3)}{V(\hat{\theta}_1)} = 1/n^2$
$\frac{V(\hat{\theta}_3)}{V(\hat{\theta}_2)} = 1/n$

5.5.1. Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from $f_Y(y;\theta) = \frac{1}{\theta}e^{-y/\theta}$, $y > 0$. Compare the Cramér-Rao lower bound for $f_Y(y;\theta)$ to the variance of the maximum likelihood estimator for θ, $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Is $\overline{Y}$ a best estimator for θ?

Ans:
Cramer-Rao bound is θ^2/n.
$\hat{\theta}$ is best estimator.

5.5.3. Suppose a random sample of size n is taken from a normal distribution with mean μ and variance σ^2, where σ^2 is known. Compare the Cramér-Rao lower bound for $f_Y(y;\mu)$ with the variance of $\hat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Is $\overline{Y}$ an efficient estimator for μ?

Ans:
Cramer-Rao bound is σ^2/n
$\hat{\mu}$ is best estimator

5.7.1. How large a sample must be taken from a normal pdf where $E(Y) = 18$ in order to guarantee that $\hat{\mu}_n = \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ has a 90% probability of lying somewhere in the interval $[16, 20]$? Assume that $\sigma = 5.0$.

Ans: 17

$\rightarrow z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 2 \rightarrow 1.645 \times \frac{5}{\sqrt{n}} = 2 \rightarrow \sqrt{n} = 4.11 \rightarrow n = 16.8$

$n \geq 17$

5.4.1) $P(|\theta - 3| > 1) = ?$

چون بدون replacement انتخاب انجام میشه

نه ترتیب مهمه نه جایگزین:

$1,2 \Rightarrow 1.5$ $2,3 \Rightarrow 2.5$ $3,4 \Rightarrow 4.5$

$1,3 \Rightarrow 2$ $2,4 \Rightarrow 3$ $3,5 \Rightarrow 4$

$1,4 \Rightarrow 2.5$ $2,5 \Rightarrow 3.5$ $4,5 \Rightarrow 4.5$

$1,5 \Rightarrow 3$

از بین 10 تا ۲ تا [illegible]

$P(|\theta - 3| > 1) = 2/10$ چون ۲ حالت هست که [illegible] اختلافش با ۳ بیشتر از ۱ است.

5.4.4) $n = 16$

$\sigma = 10$

if $\mu = 20 \rightarrow P(19 < \hat{\mu} < 21) = ?$

$\hat{\mu} = \bar{Y}$

چون $n > 15$ پس sample نرمال [illegible]

~~$\rightarrow$ [illegible] σ^2 / n~~ $\rightarrow \frac{100}{16}$

$$P\left(\frac{19-20}{10/\sqrt{16}} < Z < \frac{21-20}{2.5}\right)$$

$$= P(-0.8 < Z < 0.8) = 0.468$$

ⓢ سوال: چه موقع $\frac{X-\mu}{\sigma}$ استفاده کنیم یا $\frac{X-\mu}{\sigma/\sqrt{n}}$

5.4.9) از اونجایی که unbiased estimator باشه، پس باید امید estimator با مقدار واقعی برابر باشه

$$E[cY_1 + 2cY_2] = 1/\theta$$

$$E y = \int_0^{1/\theta} 2y^2\theta^2\, dy = \frac{2}{3\theta}$$

$$\rightarrow 3c\, E y = 1/\theta \rightarrow 3c\left(\frac{2}{3\theta}\right) = 1/\theta \rightarrow \theta = 1/2$$

5.4.11)

اگر فرض کنیم unbiased هست:

$E(\theta) = W$

$E(\theta^2) = V(\theta) + (E(\theta))^2 \rightarrow W^2$

چنین حالتی unbiased نیست. مگر اینکه بگیم consistant هست که در بی‌نهایت [illegible] n بزرگ، [illegible] estimate [illegible]. [illegible] $V(\theta) = 0$ [illegible].

پس در این حالت [illegible]، [illegible]

The principal randomly selected six students to take an aptitude test. Their scores were: (assume $X \sim$ Normal)

89.8 | 76.3 | 79.4 | 87.6 | 79.4 | 70.9

Determine a 90% confidence interval for the mean score for all students.

- Small sample, unknown μ and σ^2

$$\left(\bar{X}_n - t_{n-1,1-\alpha/2}\sqrt{\frac{\hat{\sigma}_n^2}{n}},\ \bar{X}_n + t_{n-1,1-\alpha/2}\sqrt{\frac{\hat{\sigma}_n^2}{n}}\right)$$

$\hat{\sigma}_n^2 = \frac{\sum(X-\bar{X})^2}{(n-1)}$?

$\bar{X}_n = 80.57$ $\quad \hat{\sigma}_n^2 = 49.8$ $\quad n=6$ $\quad \alpha = 0.1$

$t_{5,0.95} = 2.01$ $\quad (1-\alpha/2)$

$$(80.57 - 2.01\cdot\sqrt{\tfrac{49.8}{6}},\ 80.57 + 2.01\cdot\sqrt{\tfrac{49.8}{6}}) \approx (74.9;\ 86.8)$$

A population is normal with a variance of 68. Suppose you wish to estimate the population mean μ. Find the sample size needed to assure with 68 percent confidence that the sample mean will not differ from the population mean by more than 3 units.

nice one.

$$\ell/2 = 3 = z_{0.84}\cdot\sqrt{\frac{68}{n}}$$

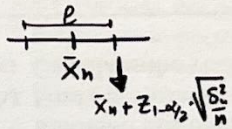
$1-\alpha = 0.68$

$\alpha = 0.32$

$\alpha/2 = 0.16$

$z_{1-\alpha/2} = z_{0.84}$

table $\rightarrow$ 1.00



Thus

$$3 = 1\cdot\sqrt{\frac{68}{n}} \rightarrow \sqrt{n} = \frac{\sqrt{68}}{3} \quad n = \frac{68}{9} = 7.5 \rightarrow n \geq 8$$

On the basis of extensive tests, the yield point of a particular type of mild steel-reinforcing bar is known to be normally distributed with $\sigma = 100$. The composition of bars has been slightly modified, but the modification is not believed to have affected either the normality or the value of σ.

a. Assuming this to be the case, if a sample of 25 modified bars resulted in a sample average yield point of 8439 lb, compute a 90% CI for the true average yield point of the modified bar.

b. How would you modify the interval in part (a) to obtain a confidence level of 92%?

a) small sample $(n=25)$, $X \sim N(\mu, \sigma^2)$, σ^2 known $(=100^2)$

$\alpha = 0.1$

$Z_{1-\alpha/2} = Z_{0.95} = 1.65$

$$I = \left[\bar{X}_n - Z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}},\ \bar{X}_n + Z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right] = 8439 \pm 1.65 \cdot \sqrt{\frac{100^2}{25}} = 8439 \pm 33 = [8406 ; 8472]$$

$t_{n-1,\ 1-\alpha/2}$

b) $\alpha = 0.08$

$Z_{1-\alpha/2} = Z_{0.96} = 1.75$

$$I = \left[\bar{X}_n - Z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}},\ \bar{X}_n + Z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right] = 8439 \pm 1.75 \cdot \sqrt{\frac{100^2}{25}} = 8439 \pm 35 = [8404 ; 8474]$$

Given the sample

5.2; 6.4; 4.8; 5.0; 5.6; **4.8**

extracted from a normally distributed population, provide an interval estimate

a) of the mean with a confidence level $\alpha = 0.05$; $\rightarrow 1-\alpha/2 = 0.975$

b) of the variance with a confidence level $\alpha = 0.1$;

$n = 6$ $\quad \bar{X}_6 = \frac{X_1 + \dots + X_6}{6} = 5.3$ $\quad S_n^2 = \frac{1}{5}\sum_{i=1}^{6}(X_i - \bar{X}_6)^2 = 0.38$

a) $I = \bar{X}_6 \pm t_{5,0.975}\sqrt{\frac{0.38}{6}} = 5.3 \pm 2.57 \cdot \sqrt{\frac{0.38}{6}} = [4.65, 5.95]$

b) $I = \left[\frac{(n-1)S_n^2}{\chi^2_{n-1,1-\alpha/2}} ; \frac{(n-1)S_n^2}{\chi^2_{n-1,\alpha/2}}\right] = \left[\frac{5 \times 0.38}{11.07} ; \frac{5 \cdot 0.38}{1.14}\right] = [0.17, 1.67]$

$\chi^2_{n-1,1-\alpha/2} \rightarrow 11.07$, $\chi^2_{n-1,\alpha/2} \rightarrow 1.14$

Given a sample of 4 measures from a population having normal ditribution, let $\bar{X}_4 = 10$ and $S_4^2 = 8$. The confidence interval of the mean, with confidence 95%, is

- ○ (a) $(10 - t_{4,0.975} \cdot \sqrt{2}, 10 + t_{4,0.975} \cdot \sqrt{2})$
- ☒ (b) $(10 - t_{3,0.975} \cdot \sqrt{2}, 10 + t_{3,0.975} \cdot \sqrt{2})$
- ○ (c) $(10 - z_{0.975} \cdot \sqrt{2}, 10 + z_{0.975} \cdot \sqrt{2})$

Given the sample $\{12, 14, 16\}$, then the confidence interval for the variance

- ○ (a) can be never provided
- ☒ (b) can be provided only if the population is normally distributed
- ○ (c) can be always provided

Given a sample of 4 measures from a population having normal ditribution, let $\overline{X}_4 = 10$ and $S_4^2 = 4$. The confidence interval of the mean, with confidence 95%, is

- ○ (a) $(10 - z_{0.975}, 10 + z_{0.975})$
- ☒ (b) $(10 - t_{3,0.975}, 10 + t_{3,0.975})$
- ○ (c) $(10 - t_{4,0.975}, 10 + t_{4,0.975})$

Let X be a population with unknown mean μ and unknown variance σ^2, not normally distributed. From a random sample of 400 observations, the point estimate of the mean and of the variance are, respectively, $\hat{\mu} = 30.5$ and $\hat{\sigma}^2 = 25$.

a) The 90% confidence interval for the mean is [30.01, 30.99]. [V] [F]
b) The 95% confidence interval for the mean is [30.01, 30.99]. [V] [F]
c) The 95% confidence interval for the variance is [20.1, 55.7] [V] [F]
d) We can provide a confidence interval for the mean because the random sample is big enough. [V] [F]
e) We can not provide a confidence interval for the mean because we do not know the true value of the variance. [V] [F]
f) We can provide a confidence interval for the variance. [V] [F]

We make 5 measures of the length of an object, collecting the following values: 30.5, 31.1, 28.6, 29.9, 29.9.

a) A point estimate of the true length, obtained considering the sample mean, is [28.6, 31.1] [V] [F]
b) The point estimate of the true length obtained considering the sample mean is 29.9 [V] [F]
c) The point estimate of the standard deviation of the measures obtained considering the unbiased sample variance is 0.829 [V] [F]
d) The point estimate of the standard deviation of the measures obtained considering the unbiased sample variance is 0.927 [V] [F]
e) Using the given sample, we can always provide an interval estimate of the true length. [V] [F]
f) Using the given sample, we can provide an interval estimate of the true length only if we know the value of the true standard deviation of the measures. [V] [F]
g) Using the given sample measures, we can provide an interval estimate of the true length only under the assumption that the measures have normally distributed random uncertainties. [V] [F]

Let X be a normally distributed population with unknown mean μ and unknown variance σ^2. We have a random sample of 3 observations, giving the values 40, 55 and 67.

a) An unbiased point estimate for the mean is 55. [V] [F]
b) An unbiased point estimate for the variance is 183. [V] [F]
c) The 90% confidence interval for the mean is [31.2, 76.8]. [V] [F]
d) The 95% confidence interval for the mean is [31.2, 76.8] [V] [F]
e) The 95% confidence interval for the variance is [74.4, 10980] [V] [F]
f) We can provide a confidence interval for the mean. [V] [F]
g) We can not provide a confidence interval for the variance, because the random sample is not big enough. [V] [F]

1. Let $X_1, \cdots, X_n$ be a sample (indipendent) drawn from a normal population. Let the variance be unknow and let us estimate from the data. The amplitude of the confidence interval (of level α) for the mean of the population

- ☒ *1* Is slightly larger than the one that could be obtained (improperly) assuming that the variance is known and equal to the value estimated from the sample.
- (b) *2* cannot be determined because there is some missing information
- (c) *3* depends on the estimated mean
- (d) *4* does not depend on the estimated variance.
- (e) *5* does not depend on the sample size.

Let $X_1, \cdots, X_n$ be an independent sample from a normal population. Let $Y_1, \cdots, Y_n$ be another sample (indipendent) of the same kind. Let the variance σ^2 of the two population be known. The amplitude of the confidence interval on the difference between the two means

- ☒ *1* Reduces as $\frac{1}{\sqrt{n}}$ as the sample size n of the two samples increases
- (b) *2* Cannot be determined because of some information is missing
- (c) *3* depends on the estimated mean
- (d) *4* grows with sample size
- (e) *5* does not depend on the sample size.

Let $X_1, \cdots, X_8$ be the result of independent measurements of the systolic blood pressure of a population with a normal diet. Let us assume it comes from a normal population. Let $Y_1, \cdots, Y_7$ another (independent) set of blood pressure measurements of patients that are following a diet with a low salt intake. Let us assume that the common variance of the two samples is unknown. Let $\bar{X} = 123.1$, $\hat{\sigma}_X = 8.4$ and $\bar{Y} = 119.3$, $\hat{\sigma}_X = 11.7$. A confidence interval for the difference $\mu_Y - \mu_X$ is

- ☒ *1* [-7.45, 15.04]
- (b) *2* [2.23, 5.82]
- (c) *3* A couple of numbers such that the probability that the parameter lies between them is fixed.
- (d) *4* A couple of numbers whose difference has the role of an estimate of the difference between the two means
- (e) *5* [-18.40, 23.56]]

$$X - Y \sim N(3.8, 3.9)$$

$$\begin{cases} \mu_{X-Y} = 123.1 - 119.3 = 3.8 \\ \sigma^2_{X-Y} = \left(\frac{8.4}{8}\right)^2 + \left(\frac{11.7}{7}\right)^2 = 1.1 + 2.8 = 3.9 \end{cases}$$

Suppose you have a simple random sample $X_1, \cdots, X_n$ from the normal distribution. You would denote by X_{n+1} a further observation (that is not available now) A prediction interval is defined as an interval in which you would expect X_{n+1} to lie with a probability 1-α. It can be found with the same method you use in confidence intervals when you have a pivot. Now what can you say about the distribution of $X_{n+1} - \bar{X}$? Suppose first that the variance is known, and then try to generalize to the case when it is unknown. Can you use this information to find a prediction interval for X_{n+1}?

Sol:

The difference $X_{n+1} - \bar{X}$ follows the Gaussian distribution with vanishing mean and a variance of $\sigma^2(1 + \frac{1}{n})$. If sigma is known we can conclude that

$$P\left(-z_{\frac{\alpha}{2}} < \frac{X_{n+1} - \bar{X}}{\sigma\sqrt{1+\frac{1}{n}}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

Inverting this region we get the following $(1-\alpha)$-prevision interval

$$\left[\bar{X} - z_{\frac{\alpha}{2}}\sigma\sqrt{1+\frac{1}{n}}, \bar{X} + z_{\frac{\alpha}{2}}\sigma\sqrt{1+\frac{1}{n}}\right]$$

In case the variance is unknown, we can proceed by combining the following two arguments.

1. Let S_n^2 be the estimator $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, then

$$(n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

2. The Student's t-distribution with n degrees of freedom can be defined as the distribution of the following ratio

$$T = \frac{Z}{\sqrt{Y/n}}$$

where Z and Y are independent. Z is a standard normal with expected value 0 and variance 1; Y has a chi-squared distribution with n degrees of freedom.

Putting the two things together we obtain

$$\frac{\frac{X_{n+1}-X}{\sigma\sqrt{1+\frac{1}{n}}}}{\sqrt{\frac{(n-1)S_n^2}{(n-1)\sigma^2}}} = \frac{X_{n+1} - \bar{X}}{S_n\sqrt{1+\frac{1}{n}}} \sim t_{n-1}$$

and therefore we can easily get that the interval

$$\left[\bar{X} - t_{\frac{\alpha}{2},n-1}\, S_n\sqrt{1+\frac{1}{n}}, \bar{X} + t_{\frac{\alpha}{2},n-1}\, S_n\sqrt{1+\frac{1}{n}}\right]$$

is a $(1-\alpha)$-prevision interval for X_{n+1}.

BERNOULLI SAMPLES

It is important that face masks used by firefighters be able to withstand high temperatures because firefighters commonly work in temperatures of 200–500°F. In a test of one type of mask, 11 of 55 masks had lenses pop out at 250°. Construct a 90% CI for the true proportion of masks of this type whose lenses would pop out at 250°.

$\hat{p} = \frac{11}{55}$

$X_i = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$ (out of 250) $\quad E[X] = p \quad V[X] = p(1-p)$

$\bar{X}_n \sim N(55p, 55p(1-p))$

$\alpha = 0.1$

$z_{1-\alpha/2} = 1.65$

$$I_{90\%} = \left(\frac{11}{55} - 1.65\sqrt{\frac{\frac{1}{5}\cdot\frac{4}{5}}{55}}, \frac{11}{55} + 1.65\sqrt{\frac{\frac{1}{5}\cdot\frac{4}{5}}{55}} \right) \simeq (0.11, 0.29)$$

A random sample of $n = 15$ heat pumps of a certain type yielded the following observations on lifetime (in years):

2.0	1.3	6.0	1.9	5.1	.4	1.0	5.3
15.7	.7	4.8	.9	12.2	5.3	.6	

a. Assume that the lifetime distribution is exponential and ~~use an argument parallel to that of Example 7.5.6~~ obtain a 95% CI for expected (true average) lifetime.

b. How should the interval of part (a) be altered to achieve a confidence level of 98%?

EXPONENTIAL SAMPLES

$$I_{1-\alpha} = \left[\frac{\chi^2_{2n,\alpha/2}}{2\sum x_i}, \frac{\chi^2_{2n,1-\alpha/2}}{2\sum x_i} \right] \quad (\text{for } \lambda)$$

$2\sum x_i = 116.8 \qquad n = 15$

(a) $\chi^2_{30,0.025} = 16.79 \qquad \chi^2_{30,0.975} = 46.98$

$$I_{95\%} = \left(\frac{16.79}{116.8}, \frac{46.98}{116.8} \right) \simeq (0.14, 0.40) \quad (\text{for } \lambda)$$

(so that $E[X]$ is in $(2.48, 6.95)$ with conf 95%)

(b) $\chi^2_{30,0.01} = 14.95 \qquad \chi^2_{30,0.99} = 50.89$

$$I_{98\%} = \left(\frac{14.95}{116.8}, \frac{50.89}{116.8} \right) \simeq (0.13, 0.44) \quad (\text{for } \lambda)$$

2: A sample of 9 measures observed from a normally distributed population assumed the following values:

0.2; 1.4; 2.3; 0.6; 2.5; −1.3; 0.8; −1.8; −0.2.

a) Provide an interval estimate of the mean, with a confidence level $\alpha = 0.1$, assuming that the variance is known and equal to 4.

b) Provide an interval estimate of the mean, with a confidence level $\alpha = 0.1$, assuming that the variance is unknown.

c) Provide an interval estimate of the variance with a confidence level $\alpha = 0.1$.

Sol: $\bar{x} = 0.5, \quad s^2 = 2.16$

a) $z_{0.95} = 1.65$, $I = [0.5 - 1.65\frac{\sqrt{4}}{\sqrt{9}}, 0.5 + 1.65\frac{\sqrt{4}}{\sqrt{9}}] = [-0.60, 1.60]$.

b) $t_{0.95} = 1.86$ (8 df), $I = [0.5 - 1.86\frac{\sqrt{2.16}}{\sqrt{9}}, 0.5 + 1.86\frac{\sqrt{2.16}}{\sqrt{9}}] = [-0.41, 1.41]$.

c) $q_{0.95} = 15.51$, $q_{0.05} = 2.73$ (8 df), $I = [\frac{8\cdot 2.16}{15.51}, \frac{8\cdot 2.16}{2.73}] = [1.11, 6.32]$.

A sample of 10 measures assumed the following values:

1.2; 0.4; 2.3; 0.6; 2.5; −1.3; 0.8; −2.8; −1.2; −0, 8.

d) Provide an interval estimate of the mean, with a confidence level $\alpha = 0.1$, assuming that the variance is known and equal to 1.

e) Provide an interval estimate of the variance with a confidence level $\alpha = 0.1$.

7.4.5. Suppose a random sample of size $n = 11$ is drawn from a normal distribution with $\mu = 15.0$. For what value of k is the following true?

$$P\left(\left|\frac{\overline{Y} - 15.0}{S/\sqrt{11}}\right| \geq k\right) = 0.05$$

Ans: $K = 2.228$

$1 - P\left(\left|\frac{\bar{Y}-15}{S/\sqrt{11}}\right| \leq k\right) = 0.95$, $1-\alpha \to 1-\alpha/2 = 0.975$, $\alpha = 0.05$

χ^2, t, $Z_{\alpha/2}$ — t-student

unknown

$t = \frac{Y-\mu}{\sigma/\sqrt{n}}$, $P(|t| \geq k) = 0.05$, $t_{10, 1-\alpha/2} = 2.228$ (0.975)

7.4.12. If a normally distributed sample of size $n = 16$ produces a 95% confidence interval for μ that ranges from 44.7 to 49.9, what are the values of $\overline{y}$ and σ?

CI: [44,7, 49.9] $t_{15, 0.975}\left(\frac{\sigma}{4}\right) = 2.6$ (2.12)

$\bar{Y} - 2.6 = 44.7$

$\bar{Y} = 46.96$ $\sigma = 4.9$

7.4.10. How long does it take to fly from Atlanta to New York's LaGuardia airport? There are many components of the time elapsed, but one of the more stable measurements is the actual in-air time. For a sample of sixty-one flights between these destinations on Sundays in April, the time in minutes (y) gave the following results:

$$\sum_{i=1}^{61} y_i = 6450 \text{ and } \sum_{i=1}^{61} y_i^2 = 684,900$$

Find a 99% confidence interval for the average flight time.

$Z_{1-\alpha/2} = 2.58 \to V(Y) = E(y^2) - (E(Y))^2 = \frac{684900}{61} - (105,73)^2 = ?$

$\bar{Y} = \frac{6450}{61} = 105.73$

To evaluate the carbon dioxide emitted from the cars we produce, a number of 50 different measures (from 50 different vehicles) are performed and recorded. Concerning this set of data, the unbiased point estimates of the mean and of the variance are, respectively, $\overline{X}_{50} = 12.5$ and $S^2_{50} = 0.36$.
a) Provide a confidence interval for the mean with confidence 99%.
b) Test, with confidence 95%, the hypothesis that the true mean is 12.0.

a) $I = \left(\bar{x}_n - z_{0,995}\sqrt{\frac{s_n^2}{n}} ; \bar{X}_n + z_{0.995}\sqrt{\frac{S_n^2}{n}}\right)$

$= \left(12.5 - 2.58\sqrt{\frac{0,36}{50}} ; 12,5 + 2.58\sqrt{\frac{0,36}{50}}\right)$

$= (12.28 ; 12.72)$

b) $Z = \frac{12.5 - 12.00}{\sqrt{\frac{s_n^2}{n}}} = 5.89$

$C = (-z_{0.975}, +z_{0.975}) = (-1.96 ; +1.96)$

$Z \notin C \Rightarrow$ reject H_0

To evaluate the carbon dioxide emitted from the cars we produce, a number of 5 different measures (from 5 different vehicles) are performed. They are listed below:

12.5; 11.1; 13.4; 11.9; 13.1.

Assume that the emissions are normally distributed.
a) Provide a confidence interval for the variance with confidence 95%.
b) Test, with confidence 95%, the hypothesis that the true variance is 0.5.

$\bar{X}_5 = \frac{12.5 + \dots + 13.1}{5} = 12.4$

$S_5^2 = \frac{(12.5-12.4)^2 + \dots + (13.1-12.4)^2}{4} = 0.86$

a) $I = \left(\frac{(n-1)S_n^2}{\chi^2_{n-1,1-\alpha/2}} , \frac{(n-1)S_n^2}{\chi^2_{n-1,\alpha/2}}\right)$ $\quad \chi^2_{4,0975} = 11.14$ $\quad \chi^2_{4,0.025} = 0.48$

$= \left(\frac{4 \cdot 0.86}{11.14} ; \frac{4 \cdot 0.86}{0.48}\right) = (0.31 ; 7.17)$

b) $V = \frac{(n-1)\cdot S_n^2}{\sigma_0^2} = \frac{4 \cdot 0.86}{0.5} = 6.88$

$C = (\chi^2_{n-1,\alpha/2}, \chi^2_{n-1,1-\alpha/2}) = (0.48 ; 11.14)$

$V \in C \Rightarrow$ Do not reject H_0

To evaluate the carbon dioxide emitted from the cars we produce, a number of 5 different measures (from 5 different vehicles) are performed. They are listed below:

12.5; 11.1; 13.4; 11.9; 13.1.

Assume that the emissions are normally distributed.
a) Provide a confidence interval for the mean with confidence 99%.
b) Test, with confidence 95%, the hypothesis that the true mean is 12.0.

$\bar{X}_5 = \frac{12.5 + \dots + 13.1}{5} = 12.4$ $\quad S_5^2 = \frac{(12.5-12.4)^2 + \dots + (13.1-12.4)^2}{4} = 0.86$

a) $I = \left(\bar{x}_n - t_{n-1,1-\alpha/2}\cdot\sqrt{\frac{S_n^2}{n}} ; \bar{X}_n + t_{n-1,1-\alpha/2}\sqrt{\frac{S_n^2}{n}}\right)$

$= \left(12.4 - 4.6\cdot\sqrt{\frac{0.86}{5}} ; 12.4 + 4.6\cdot\sqrt{\frac{0.86}{5}}\right)$ $\quad t_{4,0.995} = 4,6$

$= (10.49 ; 14.41)$

b) $T = \frac{\bar{X}_n - \mu_0}{\sqrt{\frac{S_n^2}{n}}} = \frac{12.4 - 12}{\sqrt{\frac{0.86}{5}}} = 0.96$

$C = (-t_{n-1,1-\alpha/2} ; +t_{n-1;1+\alpha/2}) = (-t_{4,0.975}, +t_{4,0.975})$

$= (-2.78 ; +2.78)$

$T \in C \Rightarrow$ Do not reject H_0

Let X be a population with unknown distribution, and unknown mean μ and variance σ^2.

a) To perform a test of hypothesis on parameter μ means to provide an interval estimate of μ. [V] [F]
b) To perform a test of hypothesis on parameter μ means to provide a point estimate of μ. [V] [F]
c) Doing an hypothesis test for the mean, with confidence 90%, we accept the null hypothesis $H_0 : \mu = 100$. However, the true value of μ can be different than 100. [V] [F]
d) Doing an hypothesis test for the mean, with confidence 90%, we accept the null hypothesis $H_0 : \mu = 100$. It means that the probability of making an error of the I kind is 0.9. [V] [F]
e) Doing an hypothesis test for the mean, with confidence 90%, we accept the null hypothesis $H_0 : \mu = 100$. It means that the probability of making an error of the I kind is 0.1. [V] [F]
f) Doing an hypothesis test for the mean, with confidence 90%, we accept the null hypothesis $H_0 : \mu = 100$. It means that the probability of making an error of the II kind is 0.9. [V] [F]
g) Doing an hypothesis test for the mean, with confidence 90%, we accept the null hypothesis $H_0 : \mu = 100$. It means that the probability of making an error of the II kind is 0.1. [V] [F]
h) The rejection region for an hypothesis test is determined by the probability of making an error of the I kind. [V] [F]
i) The rejection region for an hypothesis test is determined by the probability of making an error of the II kind. [V] [F]
j) Knowing the probability of making an error of the I kind one can also calculate the probability of making an error of the II kind. [V] [F]
k) Doing an hypothesis test for a parameter, with fixed values of confidence and dimension of the sample, if one decreases probability of making an error of the I kind restricting the rejection region, then probability of making an error of the II kind increases. [V] [F]
l) Doing an hypothesis test for a mean, using the sample mean and with fixed rejection region, if one increases the dimension n of the sample then both the probabilities of making errors of the I and of the II kind decreases. [V] [F]

Given the sample

$$0.15;\ \ 1.25;\ \ -0.55;\ \ 0.85;\ \ 2.00;\ \ -1.30;$$

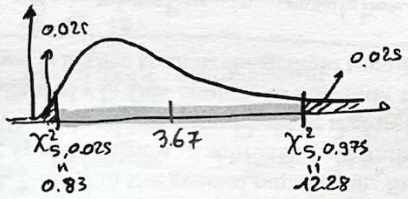
extracted from a normally distributed population, test the hypothesis $\mu = 0$ versus $\mu \neq 0$

a) Test the hypothesis that the variance is equal 2 with a confidence level $\alpha = 0.05$.

b) Testing the hypothesis that the mean is equal 0 we fix the interval $I = [-1, 1]$ as the acceptance region for the test (and the sample mean as the test statistic). Given the dimension $n = 10$ for the sample, and assuming that the variance is known and equal to 2, what it is the probability of an error of the I kind?

a) To test variance we must use $V = \frac{(n-1)S_n^2}{\sigma_0} = \frac{5 \times 1.47}{2} = 3.67$ ($H_0: \sigma^2 = 2$)

The acceptance region is determined by a χ^2_{n-1} ($n-1 = 5$ degrees freedom)

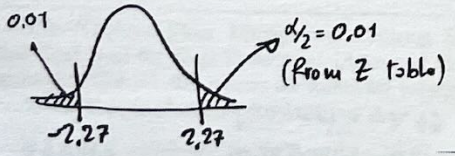


Since $V \in$ acept. the we do not reject $H_0: \sigma^2 = 2$

b) Here acceptance region = $[-1, 1]$ (for $\bar{X}_n$)

That is, we accept if $\bar{X}_n \in [-1, 1]$

which means if $Z = \frac{\bar{X}_n - 0}{\sqrt{2/n}} \in \left[\frac{-1}{\sqrt{2/10}}, \frac{+1}{\sqrt{2/10}}\right] = [-2.27, +2.27]$



Thus $\alpha = 0.02$

Thus confidence 0.98

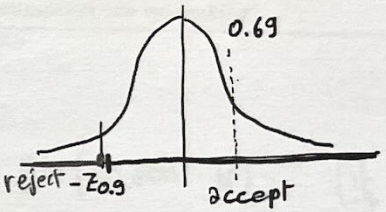
Natural cork in wine bottles is subject to deterioration, and as a result wine in such bottles may experience contamination. The article "Effects of Bottle Closure Type on Consumer Perceptions of Wine Quality" (*Amer. J. of Enology and Viticulture*, 2007: 182–191) reported that, in a tasting of commercial chardonnays, 16 of 91 bottles were considered spoiled to some extent by cork-associated characteristics. Does this data provide strong evidence for concluding that more than 15% of all such bottles are contaminated in this way? Let's carry out a test of hypotheses using a significance level of .10.

$X_i = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$ (if spoiled)

$\bar{X}_n \sim N\left(p, \frac{p(1-p)}{n}\right)$

$H_0: p \geq 0.15 \qquad H_1: p < 0.15$

$\bar{X}_n = \frac{\frac{16}{91} - 0.15}{\sqrt{\frac{0.15(1-0.15)}{91}}} \simeq 0.69$



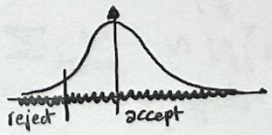
We can not reject $p \geq 15$ (more than 15% are contaminated)

Let the price paid for a given common good by different public administration be approximately normal with a mean of 34.30 euro. Suppose that the same good has been bought 26 times by Politecnico di Torino in the past year, at prices with an average of 39.54 euros and with a standard deviation of 5.2 euros. We may suspect that the price paid by our University is anomalously high and hides some misbehavior by the administration. Is there enough evidence (with a confidence level of 0.05) to exclude that such an higher mean is only due to random fluctuations? Could we keep the same decision if the requested confidence level was 0.01.

How do we compute the P-value?

Note: this is a one-sided test..

$H_0: \mu_{POLITO} > 34.30$



$n = 26$ $\quad \bar{X}_n = 39.54$ $\quad \hat{\sigma}_n = 5.2$ $\quad \hat{\sigma}_n^2 = 5.2^2$

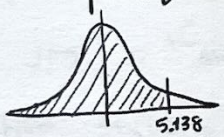
$$T = \frac{\bar{X}_n - \mu_0}{\sqrt{\hat{\sigma}_n^2 / n}} = 5.138 \qquad \text{with } 25 \text{ d.f.}$$

$\alpha = 0.05 \longrightarrow t_{25,0.05} = -1.71 \quad C = (-1.71, +\infty) \longrightarrow H_0$ not reject

$\alpha = 0.01 \longrightarrow t_{25,0.01} = -2.49 \quad C = (-2.49, +\infty) \longrightarrow H_0$ not reject

p-value

$p = P[T < 5.138] = 0.999987$



$p > \alpha$ (in both cases)

$\longrightarrow$ We do not reject (both cases)

come the fuck on!!!

The prevalence of a disease in a population is the probability that a randomly chosen individual in the population is affected by that disease. Let us focus on a specific disease whose prevalence is $2,7 \cdot 10^{-4}$ in the general population. Suppose that we want to check whether in a specific region, a source of pollutants may raise the prevalence of the disease. We pick all inhabitants of the region (42500 individuals) and record 14 people affected. Is this enough to conclude that in the region the prevalence of the disease is higher than in the general population?

Let $p_0 = 2.7 \cdot 10^{-4}$ be the proportion in population (whole)

p_1 = proportion in the specific region

$$X_i = \begin{cases} 1 \rightsquigarrow \text{prob } p \\ 0 \rightsquigarrow \text{prob } 1-p \end{cases}$$

$H_0: p_1 = p_0$ but unilateral test

(we reject if $\hat{p}_1$ is too "big")

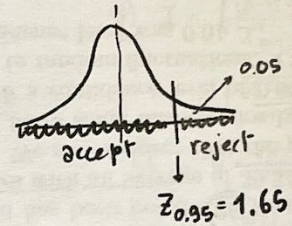
Use $\alpha = 0.05$ $\quad n = 42\,500$

① By CLT, since n is large

$$S_n = \sum_{i}^{42500} X_i \sim N(n \cdot p_0, n p_0 (1-p_0)) \quad (\text{if } p_1 = p_0)$$

so that $Z = \dfrac{S_n - E[S_n]}{\sqrt{V[S_n]}} \sim N(0,1)$

The observed Z is $Z_{obs} = \dfrac{14 - E[S_n]}{\sqrt{V[S_n]}} = 0.74$



$Z_{obs} \in$ accept region

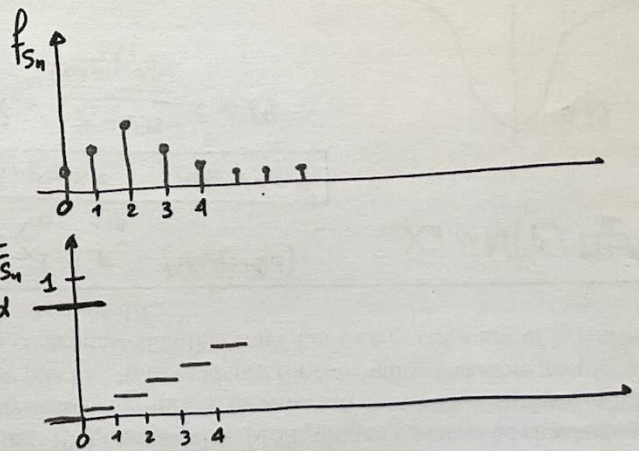
$\Downarrow$

We do not reject

(Note: $Z > 1.65 \Rightarrow \dfrac{S_n - E[S_n]}{\sqrt{V[S_n]}} > 1.65 \Rightarrow S_n > 17.06$)

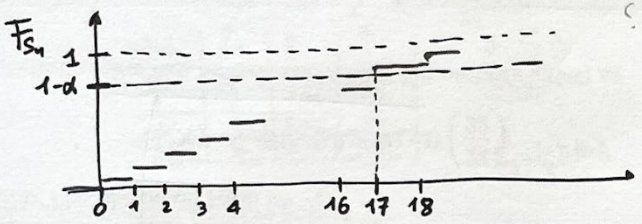
② Actually, there is no need of CLT.

In fact, $S_n \sim \text{Bin}(42'500, p_0)$ (if H_0 true)



$\alpha = P[\text{error I kind}] = P[S_n > c \mid p = p_0]$

thus c is the quantile of level $1-\alpha$ from a $Bin(42500, p_0)$



Such a c is 17 (long calculations)

Thus, for S_n → acceptance region $[0, 16]$
→ rejection region $[17, +\infty)$

We observed 14. ⇒ We do not reject.

$P_0 = \frac{82}{150}$ $\quad \alpha = 0.01 \to C: (-2.58, 2.58)$ $\quad \alpha = 0.05 \to C: (-1.96, 1.96)$

$N\left(p, \frac{p(1-p)}{n}\right) = (0.56, 0.0016) \Rightarrow Z_n = \frac{0.56 - 0.4}{0.04} = 4$

A random sample of 150 recent donations at a certain blood bank reveals that 82 were type A blood. Does this suggest that the actual percentage of type A donations differs from 40%, the percentage of the population having type A blood? Carry out a test of the appropriate hypotheses using a significance level of .01. Would your conclusion have been different if a significance level of .05 had been used?

Ans:
$Z = 3.67 \geq 2.58$
→ reject H_0
$H_0 : p = 40/100$
$H_1 : p \neq 4$

A pharmaceutical company produces a drug in pills that are expected to contain an average of 20mg of a specific active ingredient. The production mechanism is subject to an accepted natural variability. Suppose that we check the composition of 25 pills and find a sample mean of 19.7 mg of the active ingredient with a standard deviation of 1.3 mg. Can we conclude that this drug does not fulfill the required standard? Which assumption are we making?

$\mu = 20$
$\bar{X}_n = 19.7$
$n = 25$
$\sigma = 1.3$

$Z_n = \frac{19.7 - 20}{1.3/5} = -1.1$

confidence 68% → $(-1, 1)$ we cannot reject

A sample of 50 lenses used in eyeglasses yields a sample mean thickness of 3.05 mm and a sample standard deviation of .34 mm. The desired true average thickness of such lenses is 3.20 mm. Does the data strongly suggest that the true average thickness of such lenses is something other than what is desired? Test using $\alpha = .05$.

Ans: $C = (-z_{0.975}, z_{975}) \quad (-1.96, 1, 96)$
$Z = -3.12 \leq -1.96$
→ reject H_0

$Z = \frac{3.05 - 3.2}{0.34/\sqrt{50}} = -3.12$

The article "Orchard Floor Management Utilizing Soil-Applied Coal Dust for Frost Protection" (*Agri. and Forest Meteorology*, 1988: 71–82) reports the following values for soil heat flux of eight plots covered with coal dust.

34.7 35.4 34.7 37.7 32.5 28.0 18.4 24.9

The mean soil heat flux for plots covered only with grass is 29.0. Assuming that the heat-flux distribution is approximately normal, does the data suggest that the coal dust is effective in increasing the mean heat flux over that for grass? Test the appropriate hypotheses using $\alpha = .05$.

$\mu = 29$, $\bar{X}_n = 30.78$

$\hat{\sigma}_u^2 = \frac{(3.92)^2 + (4.62)^2 + (3.92)^2 + (6.92)^2 + (1.72)^2 + (2.78)^2 + (12.38)^2 + (5.88)^2}{7}$

$\hat{\sigma}_u^2 = 42.64$

$Z = \frac{30.78 - 29}{\sqrt{\frac{42.64}{8}}} = 0.773$

→ we can not reject.

acc area $= (-z_{0.025}, z_{0.975}) = (-1.96, 1.96)$

Sea water is considered suitable for swimming if it contains a pollutant with a concentration that does not exceed 0.034 mg/l. Suppose we can consider this concentration as a Gaussian random value with a know variance, equal to $4 \cdot 10^{-4}$ mg^2/l^2. In a give day the municipality check the concentration of the pollutant in 10 independent experiments and get a mean of 0.041 mg/l. At a level of confidence of 0.05, should an alarm be raised, or this value is compatible with the natural variability of the concentration of the pollutant? Which is the p-value of the test?

$Z_n = \frac{0.041 - 0.034}{\sqrt{\frac{4 \times 10^{-4}}{10}}} = 1.16$ → we cannot rejet

$\alpha = 0.05 \to C: (-z_{1-\alpha}, z_{1-\alpha})$
$(-1.645, 1.645)$

that is equivalent to

$$S_4 = \sum_{i=1}^{4} X_i > c'$$

The sum of 4 independent and identically distributed Poisson variables is Poisson with mean 4λ. The critical value c' can be fixed by asking that the level is 0.05. The level is indeed

$$P_{24}(S_n > c').$$

Under H_0, S_n is Poisson with mean 94. We therefore have to calculate which is the smallest c' such that

$$P_{24}(S_n > c') = 1 - F_{S_n}(c') \leq 0.05.$$

It turns out that $c' = 112$, and we can reject the null Hypothesis that was professor A to run the exam session.

A pair of professors share the task of examining the students for a given topic. The number of students examined in an entire day by professor A is Poisson distributed with mean 24. Professor B examines students with a rate of 32 per days. Develop a likelihood ratio test that is able to reject or not the null hypothesis that an exam session, 4 days long, where 119 students have been examined, has been run by professor A with a given confidence level.

Solution The likelihood ratio is

$$\lambda(\mathbf{X}) = \exp\left[(-24+32)\cdot 4\right]\left(\frac{24}{32}\right)^{\sum_i X_i}$$

the rejection region can be written (taking log on both sides) as

$$4(-24+32) + \sum_{i=1}^{4} X_i \log\frac{24}{32} < c$$

6.5.1. Let $k_1, k_2, \ldots, k_n$ be a random sample from the geometric probability function

$$p_X(k; p) = (1-p)^{k-1}p, \quad k = 1, 2, \ldots$$

Find λ, the generalized likelihood ratio for testing $H_0: p = p_0$ versus $H_1: p \neq p_0$.

$L_{p_0}(p|\bar{x}) = p_0^n (1-p_0)^{\sum x_i - n}$ [sample : $\bar{X} = (x_1, \ldots, x_n)$]

$\sup_p L_p(p|\bar{x}) = \left(\frac{n}{\sum x_i}\right)^n \cdot \left(1 - \frac{n}{\sum x_i}\right)^{\sum x_i - n}$

Thus

$$\lambda = \frac{p_0^n (1-p_0)^{\sum x_i - n}}{\left(\frac{n}{\sum x_i}\right)^n \cdot \left(1 - \frac{n}{\sum x_i}\right)^{\sum x_i - n}}$$

The manufacturer of a particular type of aluminum bars states that they present a number of defects, for each meter, that has Poisson distribution with mean $\lambda = 1$. To test his assertion, we check 20 piece whose length is 1 meter, obtaining the following: 0 flaws in 5 cases, 1 flaw in 7 cases, 2 flaws in 3 cases, more than 2 flaws in the remaining 5 cases. Test, with the χ^2 method and confidence level 95%, if what he affirms can not be rejected.

value	f_{oi}	p_i	f_{ti}
0	5	0,36	7,2
1	7	0,36	7,2
2	3	0,18	3,6
>2	5	0,10	2,0

$p_0 = \frac{1^0}{0!}e^{-1} =$

$p_1 = \frac{1^1}{1!}e^{-1} =$

$\vdots$

$p_{>2} = 1 - p_0 - p_1 - p_2$

$$W = \sum_i \frac{(f_{oi} - f_{ti})^2}{f_{ti}} = 5{,}28$$

$\chi^2_{N-m-1} \rightarrow \chi^2_{4-0-1} = \chi^2_3$ $\quad q_{0.95} = 7.8 \rightarrow \chi^2_{3,0.95}$

$W < q_{0.95} \rightarrow$ accept H_0

Let X be a population with unknown distribution, and unknown mean μ and variance σ^2.

n) To perform a test of hypothesis with $H_0 : X \sim N(\mu, \sigma^2)$ we can use the χ^2 test. [V] [F]

o) To perform a test of hypothesis with $H_0 : X \sim Exp(\lambda)$ we can use the χ^2 test. [V] [F]

p) Performing a test of hypothesis on the distribution of X using the χ^2 test, the number of degrees of freedom the boundary $\chi_{(1-\alpha)}$ of the acceptance region depends on the number of classes used to collect the observed data. [V] [F]

q) Performing a test of hypothesis on the distribution of X using the χ^2 test, the number of degrees of freedom of $\chi_{(1-\alpha)}$ depends on the dimension n of the sample. [V] [F]

r) Performing a test of hypothesis on the distribution of X using the χ^2 test, we need to know in advance the value of μ. [V] [F]

The following data relate the mother's age and the birthweight (in grams) of her child.

		Birthweight	
		0	1
	Maternal Age	Less Than 2,500 Grams	More Than 2,500 Grams
A	20 years or less	10	40
B	Greater than 20	15	135

Test the hypothesis that the baby's birthweight is independent of the mother's age.

$\alpha = 0.01$

X \ Y	0	1	
A	10	40	50
B	15	135	150
	25	175	200

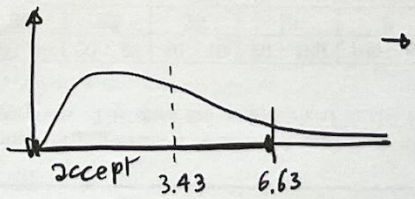
of data → co-probability data

X \ Y	0	1	
A	0.05	0.2	0.25
B	0.075	0.675	0.75
	0.125	0.875	1

$$W = \left[\frac{(10 - 0.25 \cdot 0.125 \cdot 200)^2}{(0.25 \cdot 0.125 \cdot 200)} + \dots - \dots \right] \tilde{=} 3.43$$

(p_i, p_j, n)

$\chi^2_{(2-1)(2-1),\,1-0.01} = \chi^2_{1,0.99} = 6.63$



→ We can not reject

We want to test with level of significance 0.05 the hypothesis that a given population is uniformly distributed on [0,10]. What is the conclusion if the ordered values from a sample of size 10 are the following?

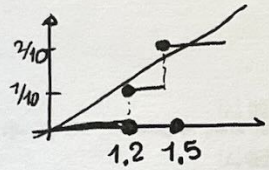
1.2 / 1.5 / 2.3 / 2.4 / 4.0 / 5.5 / 6.1 / 7.8 / 8.1 / 9.5

Here use K-S test for GOF. $F(x) = \frac{x}{10}$ $(x \in [0,10])$

For any x_i we consider 2 distances...

$d_i^+ = F(x_i) - \hat{F}(x_i^-)$

$d_i^- = F(x_i) - \hat{F}(x_i^+)$



$d_1^+ = 0.12 - 0 = 0.12$ $\quad d_1^- = 0.12 - 0.1 = 0.02$

$d_2^+ = 0.15 - 0.1 = 0.05$ $\quad d_2^- = 0.15 - 0.2 = -0.05$

$d_3^+ = 0.23 - 0.2 = 0.03$ $\quad d_3^- = -0.07$

$d_4^+ = 0.24 - 0.3 = -0.06$ $\quad d_4^- = -0.16$

$d_5^+ = 0.4 - 0.4 = 0$ $\quad d_5^- = -0.1$

$d_6^+ = 0.55 - 0.5 = 0.05$ $\quad d_6^- = -0.05$

$d_7^+ = 0.61 - 0.6 = 0.01$ $\quad d_7^- = -0.09$

$d_8^+ = 0.78 - 0.7 = 0.08$ $\quad d_8^- = -0.02$

$d_9^+ = 0.81 - 0.8 = 0.01$ $\quad d_9^- = -0.09$

$d_{10}^+ = 0.95 - 0.9 = 0.05$ $\quad d_{10}^- = -0.05$

$D = \max\{|d_i^+|, |d_i^-|\} = 0.16$

$D_{10,0.95} = 0.41$ $\quad C = (0.41, 1]$

$D \in \bar{C} \Rightarrow$ We can not reject H_0

A sample of 120 items is considered, and their lifetimes (in days) have been recorded for statistical purposes. The data are summarized in the following table:

0 - 10	10 - 20	20 - 40	40 - 70	70 - 100	100 - ∞
24	29	22	20	15	8

Test with the χ^2 method, and confidence $\alpha = 0.01$, the hypothesis that the lifetimes have exponential distribution with mean $\mu = 20$.

Ans: reject

جواب صفحه 29

---

2. According to the Mendelian theory of genetics, a certain garden pea plant should produce either white, pink, or red flowers, with respective probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$. To test this theory, a sample of 564 peas was studied with the result that 141 produced white, 291 produced pink, and 132 produced red flowers. Using the chi-square approximation, what conclusion would be drawn at the 5 percent level of significance?

$\alpha = 0.05$

جواب صفحه 29

---

3. A sample of size 120 had a sample mean of 100 and a sample standard deviation of 15. Of these 120 data values, 3 were less than 70; 18 were between 70 and 85; 30 were between 85 and 100; 35 were between 100 and 115; 32 were between 115 and 130; and 2 were greater than 130. Test the hypothesis that the sample distribution was normal.

---

4. An experiment designed to study the relationship between hypertension and cigarette smoking yielded the following data.

	Nonsmoker	Moderate Smoker	Heavy Smoker
Hypertension	20	38	28
No hypertension	50	27	18

Test the hypothesis that whether or not an individual has hypertension is independent of how much that person smokes.

Data are said to be from a *lognormal* distribution with parameters μ and σ if the natural logarithms of the data are normally distributed with mean μ and standard deviation σ. Use the Kolmogorov–Smirnov test with significance level .05 to decide whether the following lifetimes (in days) of a sample of cancer-bearing mice that have been treated with a certain cancer therapy might come from a lognormal distribution with parameters $\mu = 3$ and $\sigma = 4$.

24, 12, 36, 40, 16, 10, 12, 30, 38, 14, 22, 18

6.5.3. Let $y_1, y_2, \ldots, y_n$ be a random sample from a normal pdf with unknown mean μ and variance 1. Find the form of the GLRT for $H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$.

Ans:

$$y = \frac{(2\pi)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu_0)^2}}{(2\pi)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\bar{y}_n)^2}} = e^{-\frac{1}{2}\left((\bar{y}_n-\mu_0)^2/(1/\sqrt{n})\right)^2}$$

base the test on $z = (\bar{y}_n - \mu_0)/(1/\sqrt{n})$

1)

c_i	f_i	p_i	np_i
0–10	24	0.39	46.8
10–20	29	0.24	28.8
20–40	22	0.23	27.6
40–70	20	0.1	12
70–100	15	0.03	3.6
100–∞	8	0.01	1.2

$\lambda = 1/20 \rightarrow F_X = 1 - e^{-x/20}$

$F_0 = 0$

$F_{10} = 0.39$

$F_{20} = 0.63$

$F_{40} = 0.86$

$F_{70} = 0.96$

$F_{100} = 0.99$

$$W = \frac{\sum (f_i - np)^2}{np} = 78.$$

$W = 87.2313$

$\chi^2_{6\,(1)(1)(0.99)} = 13.27$

m n 1−α

W

13.77

we reject

2.

c_i	p_i	np_i	f_i	$\frac{(f_i-np)^2}{np}$
white	1/4	141	141	0
pink	1/2	282	291	0.28
Red	1/4	141	132	0.57

$W = 0.85$

$\chi^2_{3-1-1,\,0.95} = 3.84$

W

3.84

we accept

4.

20	38	28	86
50	27	18	95
70	65	46	181

$\div \frac{1}{181}$

0.11	0.211	0.154	0.475
0.276	0.15	0.099	0.525
0.386	0.361	0.253	1

$$W = \left[\frac{20 - 0.475 \times 0.386 \times 181}{0.475 \times 0.386 \times 181} + \frac{50 - 0.525 \times 0.386 \times 181}{0.525 \times 0.386 \times 181} + \cdots\right] =$$

$\chi^2_{(3-1)(2-1),\,0.95} = 5.991$

5.991

use $\alpha = 0.05$ for independency check

If $W > 5.991 \rightarrow$ reject

If $W < 5.991 \Rightarrow$ accept