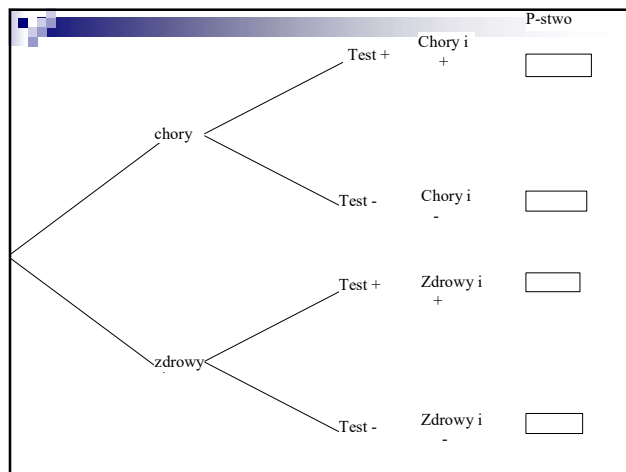


Wykład 3

Wzór Bayesa – wpływ rozkładu a priori.

- ✓ 30% ludzi jest chorych na pewną chorobę.
- ✓ Test do wykrywania tej choroby ma czułość 99.7% i specyficzność 98.5%.
- ✓ Jakie jest prawdopodobieństwo, że osoba u której test wykrył chorobę jest faktycznie chora ?



$$P(\text{chory} | \text{test}+) = \frac{P(\text{chory i test}+)}{P(\text{test}+)} =$$

P-stwo, że osoba z dodatnim (wskazującym na obecność choroby) wynikiem testu jest faktycznie chora wynosi

Zmienna losowe

- Zmienna liczbowa, której wartość zależy od wyniku eksperymentu.
- Przykład 1 (liczba orłów uzyskanych w jednym rzucie monetą)
- Przykład 2 (liczba oczek uzyskanych w rzucie kostką)

Zmienna losowa dyskretna

- Zbiór wartości, które może przyjąć zmienna losowa dyskretna jest skończony lub przeliczalny. Możliwe do uzyskania wartości będziemy oznaczali $x_1, x_2, \dots$
- Rozkład p-stwa zmiennej X określamy przyporządkowując p-stwa poszczególnym wartościom.
- Np. w rzucie symetryczną kostką liczba oczek X ma rozkład

Ciągła zmienna losowa

- Ciągła zmienna losowa może przybierać wszystkie wartości z odcinka lub całej prostej rzeczywistej.
- Rozkład zmiennej ciągłej można opisać podając jej funkcję gęstości $f(x)$.

Wartość oczekiwana i wariancja

- Zmienna losowa dyskretna
 - $E(X) = \sum x_i P(X = x_i) = \mu_x$
 - $\text{Var}(X) = \sum (x_i - \mu_x)^2 P(X = x_i) = \sum x_i^2 P(X = x_i) - \mu_x^2$
- Przykład 1 (rzut monetą)
- Przykład 2 (rzut kostką)

- Ciągła zmienna losowa

$$EX = \int_{-\infty}^{\infty} x f(x) dx$$

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - EX)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - (EX)^2$$

Własności wartości oczekiwanej i wariancji

- $E(aX+b)=aEX+b$
- $\text{Var}(aX+b)=a^2\text{Var}(X)$

Dla dwóch zmiennych losowych X i Y

- ✓ $E(X+Y)=EX+EY$
- ✓ $E(X-Y)=EX-EY$
- ✓ $E(aX+bY+c)=$
- ✓ $\text{Var}(X+Y)=\text{Var}(X)+\text{Var}(Y)$, gdy X & Y są niezależne
- ✓ $\text{Var}(X-Y)=\text{Var}(X)+\text{Var}(Y)$, gdy X & Y są niezależne

Trzy monety (Rozkład Bernoulliego)

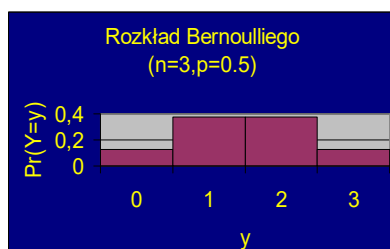
Anita, Beata i Krystyna rzucają monetami.

A	B	K	P-stwo
O	O	O	
O	O	R	
O	R	O	
O	R	R	
R	O	O	
R	O	R	
R	R	O	
R	R	R	

Zdarzenie	P-stwo
0R 3O	
1R 2O	
2R 1O	
3R 0O	

Zmienna losowa Y = liczba orłów
Podaj rozkład zmiennej Y

- Histogram w populacji. Populacja – ``wszystkie’’ rzuty trzema monetami



Ciąg prób Bernoulliego

- n niezależnych powtórzeń tego samego eksperymentu
- Dwa możliwe wyniki w każdej próbie - ``sukces’’ lub ``porażka’’ (np. O lub R, albo 0 lub 1)
- W każdej próbie p -stwo sukcesu wynosi p
- Y = liczba sukcesów w n próbach (np. liczba orłów, liczba heterozygot)

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$= \frac{n(n-1)\dots(n-k+1)}{k(k-1)\dots 1}$$

$$\binom{n}{y}$$

- Symbol Newtona
- Liczba możliwych ``ustawień’’ y sukcesów i $n-y$ porażek

Przykład z monetami.

$$\binom{3}{0} =$$

$$\binom{3}{1} =$$

$$\binom{3}{2} =$$

$$\binom{3}{3} = ; n =$$

$p =$

- $P(Y=0) =$

- $P(Y=1) =$

- $P(Y=2) =$

- $P(Y=3) =$

- Rozkład jest symetryczny bo $p=1/2$.

Efekt uboczny lektarstwa

- 20% ludzi dostaje nudności po zażyciu pewnego lektarstwa
- Lekarz przepisał lektarstwo czterem nowym pacjentom
- Y – liczba pacjentów w naszej próbie, którzy dostali nudności
- Podaj rozkład zmiennej Y

Odpowiedź

- $P(\text{co najmniej 2 dostanie nudności}) =$
- $P(\text{przynajmniej 1 dostanie nudności}) =$

Parametry rozkładu Bernoulliego

- $EY = np$
- $\text{Var } Y = np(1-p)$

Niektóre własności symbolu Newtona

- $\binom{n}{0} = \binom{n}{n} =$
- $\binom{n}{1} = \binom{n}{n-1} =$
- Ogólnie $\binom{n}{y} = \binom{n}{n-y}$

Przykład

- Jeden na ośmiu dorosłych mężczyzn ma podniesiony poziom cholesterolu. Losowo wybrano 6 mężczyzn z populacji. Jakie jest p-stwo, że (dokładnie) 2 spośród nich ma podniesiony poziom cholesterolu ?

- Jakie jest p-stwo, że co najmniej jeden z nich ma podniesiony poziom cholesterolu ?

Dystrybuanta

- $F_Z(x) = P(Z \leq x)$ – dystrybuanta zmiennej losowej Z w punkcie x
- Własności:
- $F_Z(x)$ jest funkcją

$$\lim_{x \rightarrow -\infty} F(x) =$$

$$\lim_{x \rightarrow \infty} F(x) =$$

Przykłady

- Narysuj dystrybuantę rozkładu
- Bernoulliego $B(3, 0.5)$
- Jednostajnego na odcinku $[a, b]$

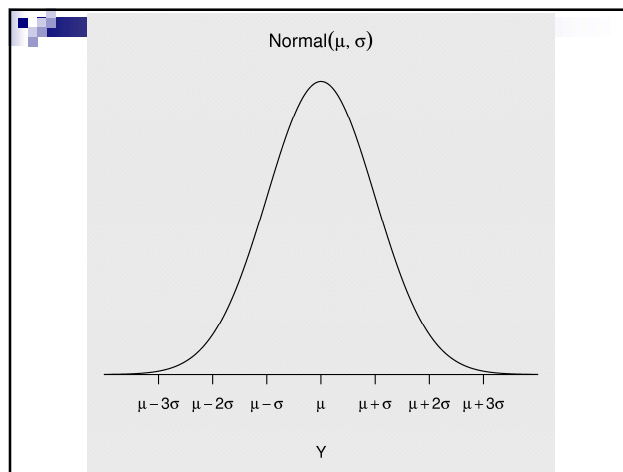
Rozkład normalny

- Bardzo często używany do modelowania rozkładów symetrycznych zmiennych losowych ciągłych
- Przykłady
 - Błąd pomiarowy
 - Wzrost, wydajność
 - Temperatura ciała
 - Zawartość różnych składników we krwi

Funkcja gęstości

- $Y \sim N(\mu, \sigma)$
- μ - wartość oczekiwana, σ - odchylenie standardowe

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$



Standardowy rozkład normalny $N(0,1)$;

- $\mu = 0, \sigma = 1$.
- Oznaczenie:
- Do oznaczenia zmiennej losowej o rozkładzie standardowym normalnym będziemy zwykle używali litery Z
- Dystrybuenta rozkładu normalnego $N(0,1)$:
- $\Phi(x) = P(Z < x)$.
- $\Phi(0) =$

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

Korzystanie z Tablic

- $P(Z < 0.95) =$
- $P(Z > 0.75) =$
- $P(Z < -1.5) =$
- $P(1.12 < Z < 2.24) =$
- $P(Z > 1.96) =$

Pożyteczne wzory

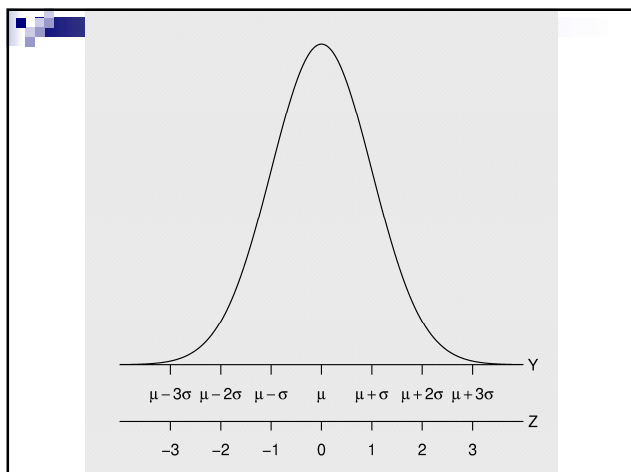
- $\Phi(-x) =$
- $P(Z > z) =$
- $P(z_1 < Z < z_2) =$
- Oblicz
 $\Pr(|Z| > 1.96) =$

Niestandardowy rozkład normalny : $N(\mu, \sigma)$

- Załóżmy, że poziomy cholesterolu w pewnej populacji mają rozkład normalny o średniej $\mu = 220$ i odchyleniu standardowym $\sigma = 40$.
- Y ma rozkład $N(220, 40)$
- Jaka część populacji ma poziom cholesterolu powyżej 240 ?

Standardyzacja

- $Y \sim N(\mu, \sigma)$
- $Z = (Y - \mu) / \sigma$ ma rozkład normalny o średniej
- $EZ =$
- i wariancji
- $\text{Var}(Z) =$



Przykład cd.

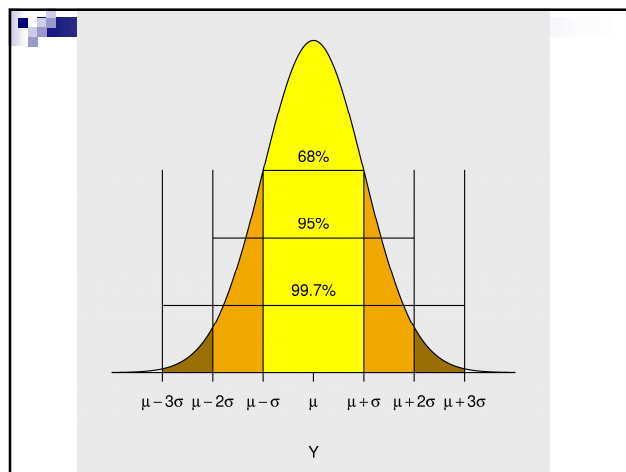
- $P(Y > 240)$?
- $P(Y > y)$ gdzie $y=240$.
- $z = (y-\mu)/\sigma = (240-220)/40 = 0.5$.
- $P(Y > 240) = P(Z > 0.5) =$

- Jakie jest p-stwo, że u losowo wybranej osoby cholesterol będzie pomiędzy 200 a 260 ?
- $y_1 = 200$; $z_1 = (200-220)/40 = -0.5$;
- $y_2 = 260$; $z_2 = (260-220)/40 = 1.0$;
- $P(200 < Y < 260) = \Pr(-0.5 < Z < 1.0) =$

- Oblicz $P(Y < 170)$

Reguła 68%–95%–99.7% (albo reguła 3σ)

- Jeżeli zmienna X ma rozkład normalny to
- $P(\mu - \sigma < X < \mu + \sigma) =$
- $P(\mu - 2\sigma < X < \mu + 2\sigma) =$
- $P(\mu - 3\sigma < X < \mu + 3\sigma) =$



Kwantyle

- W jakim punkcie y dystrybucja osiąga zadaną wartość p ?
- Przykład : Jaki jest trzeci kwantyl, tzn. kwantyl rzędu ?

- Znajdź trzeci kwantyl rozkładu opisującego poziom cholesterolu.

- Znajdź kwantyl rzędu 0.1 dla rozkładu poziomów cholesterolu.

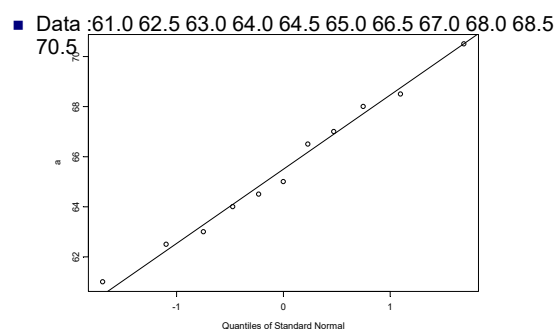
Ocena normalności

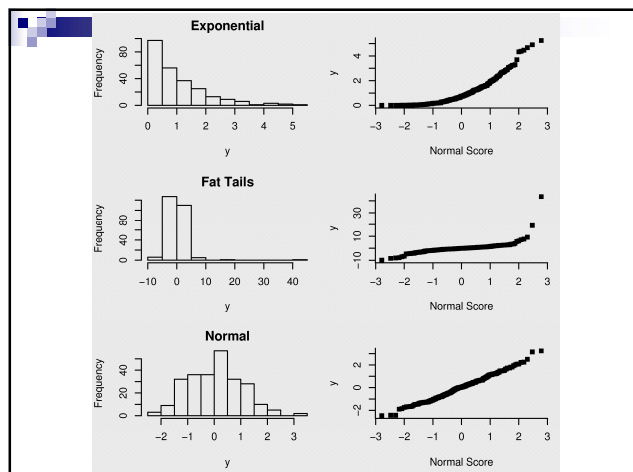
- Znaczna część procedur statystycznych, które poznamy w dalszej części kursu wymaga założenia, że próba pochodzi z populacji o rozkładzie normalnym. Założenie to można sprawdzać np. wykonując proste obliczenia lub rysując wykres kwantyl-kwantyl.

Reguła 3 σ

- Policzmy procent obserwacji, które znajdują się w odległości $\pm 1s$, $\pm 2s$ and $\pm 3s$ od $\bar{y}$.
- Przykład: poziomy serum CK
- $n = 36$, $\bar{y} = 98.28$ i $s = 40.38$.
- $26/36 = 72\%$ obserwacji jest w przedziale $\bar{y} \pm 1s$
- $34/36 = 94\%$ obserwacji jest w przedziale $\bar{y} \pm 2s$
- $36/36 = 100\%$ obserwacji jest w przedziale $\bar{y} \pm 3s$
- To w przybliżeniu odpowiada wartościom dla rozkładu normalnego.

Wykres kwantyl-kwantyl (QQ plot)





Korekta na ciągłość

- Zwykle używamy rozkładu normalnego dla danych, które nie są ciągłe.
- Dyskretyzacja pomiarów
- Niektóre rozkłady dyskretne można dobrze przybliżać rozkładem normalnym
- Gdy zbiór możliwych do uzyskania wartości jest duży i duży jest rozmiar próby to możemy zaniedbać korektę na ciągłość.
- Korektę na ciągłość często stosujemy gdy przybliżamy rozkład Bernoulliego rozkładem normalnym.

Przykład

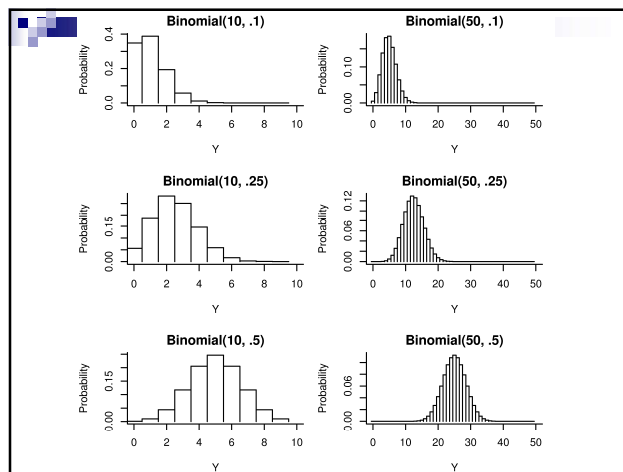
- Wyniki testu
 - ☐ Załóżmy, że $\mu = 100$ i $\sigma = 16$.
 - ☐ Wynik to liczba całkowita – nie pochodzi z rozkładu ciągłego.
 - ☐ Jak użyć rozkładu normalnego ?
 - ☐ Przypisujemy liczbie całkowitej y p-stwo z odcinka od $y-.5$ do $y+.5$

- ☐ Jakie jest p-stwo, że losowo wybrany student uzyskał wynik między 120 a 140 punktów.

Przybliżenie rozkładu Bernoulliego rozkładem normalnym

Po co ?

- Gdy n jest duże mamy bardzo dużo możliwych wartości Y
- Gdy n jest duże symbole Newtona są trudne do wyliczenia
- Działa tym lepiej im p jest bliższe 0.5



- Rozkład Bernoulliego z parametrami n i p przybliżamy rozkładem normalnym z $\mu=np$ i $\sigma=\sqrt{np(1-p)}$.
- Przybliżenie jest "dość dobre" gdy $np \geq 5$ i $n(1-p) \geq 5$
- Przybliżenie jest dość dokładne w centrum rozkładu i gorsze w "ogonach".

Przykład

- Załóżmy, że Y ma rozkład Bernoulliego z $n=40$ i $p=0.25$. Wtedy $\mu = np =$ i $\sigma =$. (sprawdzamy $n(1-p) =$).
- Chcemy policzyć p-stwo, że Y jest pomiędzy 10 i 15 (włącznie).
- $P(10 \leq Y \leq 15) =$
- [Wynik dokładny = 0.534; nie najgorzej.]

Rozkłady próbkowe

- Rozważmy populację o pewnym rozkładzie, np. normalnym lub dwupunktowym ($P(Y=1)=p$, $P(Y=0)=1-p$).
- Parametry - μ i σ , lub p-stwo sukcesu p .
- Bierzemy próbę o rozmiarze n z populacji, lub wykonujemy n prób Bernoulliego.
- próba $y_1, \dots, y_n$, lub y = liczba sukcesów.
- liczymy $\bar{y}$ i s , lub $\hat{p}$ = (estymatory)
- Gdy rozmiar próby jest duży to estymatory są na ogół bliskie parametrom które estymują.

- Jak bardzo estymatory mogą się różnić od prawdziwych parametrów ?
- Co się stanie jeżeli weźmiemy inną próbę?
- Będziemy mieli inne wartości $\bar{y}$ i $\hat{p}$.
- Interesuje nas rozkład próbkowy $\bar{y}$ i $\hat{p}$.

Meta-eksperyment

- Wobraźmy sobie, że powtarzamy eksperyment wiele razy
- Interesuje nas rozkład p-stwa wszystkich możliwych do uzyskania wartości $\bar{y}$, s lub $\hat{p}$
- Taki rozkład będziemy nazywali rozkładem próbkowym estymatora.
- Zwykle próbujemy tylko raz.
- Rozkłady próbkowe można obliczyć teoretycznie.

Rozkład próbkowy estymatora p w rozkładzie dwupunktowym

- Y = liczba sukcesów w n próbach
- y = zaobserwowana liczba sukcesów
- $\hat{p} = \frac{y}{n}$ - estymator p

Przykład:

- Producent ocenia, że 2% jego wyrobów jest wadliwych. Wyroby te paczkuje się po 40 w jednym opakowaniu.
- Y = liczba wadliwych wyrobów w losowo wybranej paczce. Y ma rozkład .
- Niech $\hat{p} = Y/40$ = frakcja elementów wadliwych.
- $\Pr(\hat{p} = \pi) =$

$$\Pr(\hat{p} = 0) = \Pr(Y = 0) = (1)(.02)^0(.98)^{40} = 0.45$$

$$\Pr(\hat{p} = 0.025) = \Pr(Y = 1) = (40)(.02)^1(.98)^{39} = 0.36$$

$$\Pr(\hat{p} = 0.05) = \Pr(Y = 2) = (780)(.02)^2(.98)^{38} = 0.14$$

$$\Pr(\hat{p} = 0.075) = \Pr(Y = 3) = (9880)(.02)^3(.98)^{37} = 0.04$$

$$\Pr(\hat{p} \geq 0.1) = \Pr(Y \geq 4) \approx 0.01$$

- Gdybyśmy otworzyli tysiące paczek to rozkład frakcji liczby wadliwych elementów w paczce byłby zgodny z rozkładem wyliczonym na poprzedniej stronie.

„Prawdziwa” wartość p jest 0.02 i nie jest możliwa do uzyskania w pojedynczym eksperymencie. Ale na ogół będziemy dostawali wartości bliskie 0.02. P-stwo, że dostaniemy $\hat{p} = 0.025$ wynosi 36%. P-stwo, że będziemy w odległości nie większej niż 0.03 od prawdziwej wartości wynosi

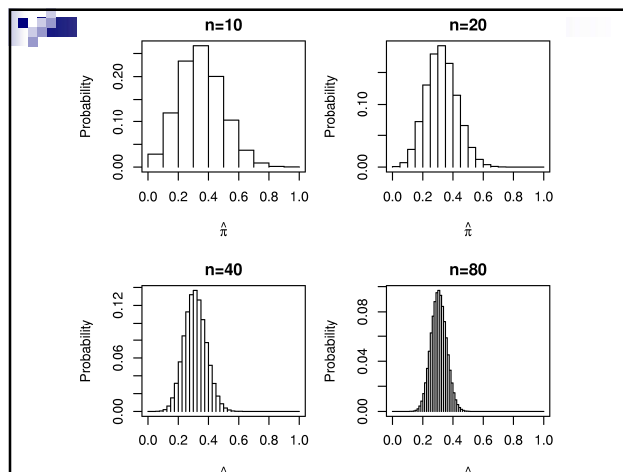
Zatem jeżeli znajdziemy 3 lub więcej wyrobów wadliwych w jednej paczce mamy podstawy, żeby kwestionować twierdzenie producenta. Sugestia – sprawdzić więcej paczek.

Przykład

- $n=40$ i $p = 0.02$. Jakie jest p-stwo, że estymator p jest co najmniej dwa razy większy od prawdziwej wartości ?

Zależność od rozmiaru próby

- Y ma rozkład Bernoulliego (n,p)
- $\mu_Y = np$
- $\text{Var}(Y) = np(1-p)$
- $\hat{p} =$
- $\mu_{\hat{p}} =$
- $\text{Var}(\hat{p}) =$
- Gdy n rośnie wariancja $\hat{p}$ i estymator staje się bardziej
- Przykład; $p=0.3$.



Rozkład $\hat{p}$ ($p=0.3$)

n	$\Pr(0.25 \leq \hat{p} \leq 0.35)$
10	0.5
20	0.535
40	0.612
80	0.728
500	0.987