Statistik för Biologer
F2: Slumpvariabler och Vanliga Fördelningar

Shaobo Jin

Matematiska institutionen

Slumpvariabel

Definition (Slumpvariabel)

En **slumpvariabel**, ofta betecknad X, är ett tal som beskriver utfallet av ett "försök" vars resultat inte är givet på förhand. Sannolikheten att X antar olika värden bestäms av dess fördelning.
Fördelningsfunktion av slumpvariabeln X definieras som $F(z) = P(X \leq z)$.

Exempel:

1. Antalet ögon som kommer upp vid kast av tärningar när vi spelar brädspelet Catan.
2. Antalet deletioner i en kromosom vid replikering.
3. Hur långt en Klebsiella-bakterie rör sig på en timme.

Sannolikhetsfunktion

Definition (Diskret variabel)

Slumpvariabeln X är **diskret** om den bara kan anta speciella värden på den skala som används - normalt bara heltal.

Diskreta slumpvariabler beskriver ofta antal.

1. Antalet ögon som kommer upp vid kast av tärningar när vi spelar brädspelet Catan.
2. Antalet deletioner i en kromosom vid replikering.

Fördelningen av en diskret variabel beskrivs av sannolikhetsfunktionen

$$p(k) = P(X = k).$$

Kullstorlek Hos Grizzlybjörn

Grizzlybjörnar får 1-5 ungar per kull. Studier har visat att sannolikheterna för olika antal ungar ser ut som följer:

Antal Ungar	Sannolikhet
1	0.11
2	0.47
3	0.40
4	0.01
5	0.01

Den sammanlagda sannolikheten är 1 eftersom något händer alltid!

- Slumpvariabel X: antalet ungar i en slumpmässigt vald kull.
- Sannolikhetsfunktionen för fördelningen för antalet ungar X:

$$p(1) = 0.11, \; p(2) = 0.47, \; p(3) = 0.40.....$$

Att Räkna med Slumpvariabel

Att Räkna med Slumpvariabel

Antal Ungar	$p(k) = P(X = k)$
1	0.11
2	0.47
3	0.40
4	0.01
5	0.01

Vad är sannolikheten att det blir två ungar i en grizzlybjörnskull?

$$P(X = 2) = 0.47.$$

Att Räkna med Slumpvariabel

Att Räkna med Slumpvariabel

Antal Ungar	$p(k) = P(X = k)$
1	0.11
2	0.47
3	0.40
4	0.01
5	0.01

Vad är sannolikheten att det blir minst tre ungar i en grizzlybjörnskull?

$$\begin{aligned} P(X \geq 3) &= P(X = 3) + P(X = 4) + P(X = 5) \\ &= 0.40 + 0.01 + 0.01 = 0.42. \end{aligned}$$

Räkneregler för Sannolikheter: Komplementhändelse

Att Räkna med Slumpvariabel

Antal Ungar	$p(k) = P(X = k)$
1	0.11
2	0.47
3	0.40
4	0.01
5	0.01

Vad är sannolikheten att det blir minst tre ungar i en grizzlybjörnskull?

- Låt $A = \{X \geq 3\}$ och $A^c = \{X < 3\} = \{X \leq 2\}$.
- Räknereglerna för sannolikheter gäller precis som tidigare. Det gäller att $P(A^c) = 1 - P(A)$.
- Sen

$$P(X \geq 3) = 1 - P(X < 3) = 1 - (0.11 + 0.47) = 0.42.$$

Population och Stickprov

- Avsikten med en statistisk undersökning är att skaffa kunskap om en stor mängd enheter. Alla enherter av intresse utgör en **population**.
- En mätning av egenskaper för en enhet kallas en **observation**.
- Samlingen av observationer kallas ett **stickprov**.

Exempel: Väljarbarometern

- Population = alla svenska.
- Stickprov = personer som har blivit intervjuade

Egenskaper Hos Slumpvariabler

Vi sammanfattar ofta informationen i ett stickprov med medelvärde och (stickprovs) varians (eller standardavvikelse). På samma sätt kan vi sammanfatta beteendet hos en slumpvariabel med hjälp av väntevärde och varians (eller standardavvikelse).

1. Väntevärdet $E(X)$ beskriver vad det genomsnittliga/"förväntade" värdet på X är.
2. Variansen $V(X)$ beskriver hur stor den genomsnittliga avvikelsen från väntevärdet är.

Väntevärde

För en diskret slumpvariabel X definieras **väntevärdet** som

$$E(X) = \sum_k k \cdot p(k).$$

där $p(k) = P(X = k)$ är sannolikhetsfunktionen för X och summan går över alla möjliga värden på k.

Antal Ungar	$p(k) = P(X = k)$
1	0.11
2	0.47
3	0.40
4	0.01
5	0.01

Väntevärdet för antalet ungar i en grizzlybjörnskull är

$$E(X) = 1 \cdot 0.11 + 2 \cdot 0.47 + 3 \cdot 0.40 + 4 \cdot 0.01 + 5 \cdot 0.01 = 2.34.$$

Varians och Standardavvikelse

För en diskret slumpvariabel X definieras **variansen** som

$$V(X) = \sum_k [k - E(X)]^2 \cdot p(k).$$

där $p(k) = P(X = k)$ är sannolikhetsfunktionen för X och summan går över alla möjliga värden på k. Den motsvarande standardavvikelsen är $S(X) = \sqrt{V(X)}$.

Varians och Standardavvikelse

$$V(X) = \sum_k [k - E(X)]^2 \cdot p(k).$$

Antal Ungar	$p(k) = P(X = k)$
1	0.11
2	0.47
3	0.40
4	0.01
5	0.01

Variansen för antalet ungar i en grizzlybjörnskull är

$$\begin{aligned} V(X) &= (1-2.34)^2 \cdot 0.11 + (2-2.34)^2 \cdot 0.47 + (3-2.34)^2 \cdot 0.40 \\ &\quad + (4-2.34)^2 \cdot 0.01 + (5-2.34)^2 \cdot 0.01. \end{aligned}$$

Population och Stickprov

1. Väntevärdet $E(X)$ beskriver vad det genomsnittliga/"förväntade" värdet på X är.
   - Om x_1, x_2, ..., x_n är våra n mätvärden ges medelvärdet av
   $$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$
   - Vi kan tänka på $E(X)$ som det medelvärdet vi skulle få om vi studera X oändligt många gånger ($n = \infty$).
   - Medelvärdet är en **skattning** av väntevärdet.
2. Variansen $V(X)$ beskriver hur stor den genomsnittliga avvikelsen från väntevärdet är.
   - Vi kan tänka på $V(X)$ som den stickprovsvariansen vi skulle få om vi studera X oändligt många gånger ($n = \infty$).
   - Stickprovsvarians
   $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
   är en **skattning** av $V(X)$.

Binomialfördelning

I många situationer upprepar man ett försök n gånger och räknar hur många gånger man får ett visst utfall:

1. Man undersöker $n = 100$ slumpmässigt utvalda individer i en population och räknar hur många som har IgG-antikroppar mot SARS-CoV-2.
2. Bläckfisken Paul tippar $n = 8$ matcher i fotbolls-VM 2010 och räknar hur många gåner han har rätt.

Binomialfördelning

Definition

Antag att vi upprepar ett försök n gånger och att det vid varje försök är sannolikhet p att händelsen A inträffar. Låt X vara antalet gånger som händelsen A inträffar. Då är X **binomialfördelad** med parametrarna n och p.

1. De möjliga utfallen för X är $k = 0, 1, 2, ..., n-1, n$.
2. Den sannolikhetsfunktion är
$$p(k) = P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k},$$
där $n! = n \cdot (n-1) \cdot \cdots \cdot 3 \cdot 2 \cdot 1$ är produkten av de n första positiva heltalen. Vi definierar $0! = 1$.
3. Kodbeteckning: $X \sim \mathrm{Bin}(n, p)$.

Binomialfördelning: Bläckfisken Paul

Exempel: Vi antar att

1. Varje gånger är sannolikheten att tippa rätt vinnare $p = 0.5$
2. Varje ny tippning är oberoende av tidigare tippningar

Vad är sannolikheten att lyckas med 8 rätt av 8 möjliga i fotbolls-VM 2010?

Antalet rätt X som Paul får är binomialfördelat med $n = 8$ och $p = 0.5$.

$$p(8) \quad = \quad P(X = 8) = \frac{8!}{8!\,(8-8)!} 0.5^8 (1 - 0.5)^{8-8} \approx 0.0039.$$

Beräkning med R:

```
dbinom(8, 8, 0.5) # (X=k, n, p)

## [1] 0.00390625
```

Binomialfördelning: Bläckfisken Paul

Exempel: Vi antar att

1. Varje gånger är sannolikheten att tippa rätt vinnare $p = 0.5$
2. Varje ny tippning är oberoende av tidigare tippningar

Vad är sannolikheten att lyckas med minst 7 matcher?

$$p(7) + p(8) = \frac{8!}{7!(8-7)!} 0.5^7 (1-0.5)^{8-7} + \frac{8!}{8!(8-8)!} 0.5^8 (1-0.5)^{8-8}$$

Med R:

```
dbinom(7, 8, 0.5) + dbinom(8, 8, 0.5) # (X=k, n, p)

## [1] 0.03515625
```

Egenskaper Hos Binomialfördelning

Om X är binomialfördelad med parametrarna n och p så är:

1. Väntevärdet $E(X) = np$
2. Variansen $V(X) = np(1-p)$
3. Standardavvikelsen $S(X) = \sqrt{V(X)} = \sqrt{np(1-p)}$.

Pauls Tippning

Med $n = 8$ och $p = 0.5$ får vi

$$\begin{aligned} E(X) &= 8 \cdot 0.5 = 4, \\ V(X) &= 8 \cdot 0.5 \cdot (1 - 0.5) = 2, \\ S(X) = \sqrt{V(X)} &= \sqrt{2}. \end{aligned}$$

Binomialfördelning

Om X är binomialfördelad är de möjliga utfallen för X $k = 0, 1, 2, ..., n-1, n$.

- X =Antalet studenter som kommer till föreläsningen idag.
- n =Antalet registerade studenter

I många situationer kan en slumpvariabel (i princip) anta hur stora värden som helst.

- Vi kan också undersöker antalet frågor som ni ställer idag. Låt X =antalet frågor. De möjliga utfallen för X är $k = 0, 1, 2, ..., n-1, n, n+1, n+2, \cdots$.

Poissonfördelning

Definition

X är **Poissonfördelad** med parametrarna m om

$$p(k) = P(X = k) = \frac{m^k}{k!} e^{-m}.$$

De möjliga utfallen för X är $k = 0, 1, 2, ...$ Kodbeteckning: $X \sim \text{Po}(m)$.

Exempler:

1. Antal bilar som passerar en vägkorsning under en timme.
2. Antal jordbävningar i Japan under ett år.
3. Antal mål i en fotbollsmatch.
4. Antal olyckor i ett lab under ett år.
5. Antal patienter som söker vård per dag.

Egenskaper Hos Poissonfördelning

Om X är Poissonfördelad med parametern m så är:

1. Väntevärdet $E(X) = m$,
2. Variansen $V(X) = m$,
3. Standardavvikelsen $S(X) = \sqrt{V(X)} = \sqrt{m}$.

Mutation

En cell utsätts för röntgenstrålning sker i genomsnitt 0.2 mutationer per dag.

$$\begin{aligned} E(X) &= 0.2, \\ V(X) &= 0.2, \\ S(X) = \sqrt{V(X)} &= \sqrt{0.2}. \end{aligned}$$

Poissonfördelning: Mutation

Mutation

En cell utsätts för röntgenstrålning sker i genomsnitt 0.2 mutationer per dag. Vad är sannolikheten att ingen mutation sker under en dags bestrålning?

X =Antalet mutationer är Poissonfördelat med $m = 0.2$.

$$p(0) = P(X = 0) = \frac{0.2^0}{0!} e^{-0.2} \approx 0.82.$$

Med R:

```
dpois(0, 0.2) # (k, m)

## [1] 0.8187308
```

Poissonfördelning: Mutation

Mutation

En cell utsätts för röntgenstrålning sker i genomsnitt 0.2 mutationer per dag. Vad är sannolikheten att mest en mutation sker under en dags bestrålning?

X =Antalet mutationer är Poissonfördelat med $m = 0.2$.

$$P(X \leq 1) = p(0) + p(1) \quad = \quad \frac{0.2^0}{0!}e^{-0.2} + \frac{0.2^1}{1!}e^{-0.2}.$$

Med R:

```
dpois(0, 0.2) + dpois(1, 0.2) # (k, m)

## [1] 0.9824769
```

Poissonfördelning: Mutation

Mutation

En cell utsätts för röntgenstrålning sker i genomsnitt 0.2 mutationer per dag. Vad är sannolikheten att minst två mutationer sker under en dags bestrålning?

X =Antalet mutationer är Poissonfördelat med $m = 0.2$.

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) = 1 - P(X \leq 1) \\ &= 1 - \left[\frac{0.2^0}{0!} e^{-0.2} + \frac{0.2^1}{1!} e^{-0.2}\right]. \end{aligned}$$

Med R:

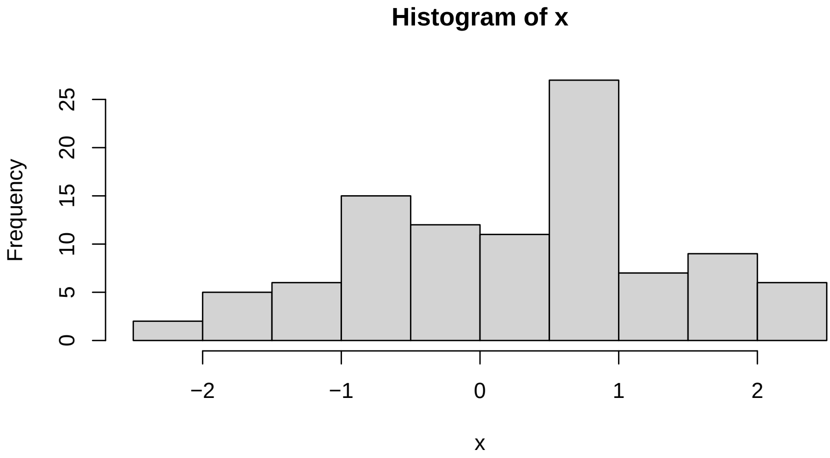
```
1 - dpois(0, 0.2) - dpois(1, 0.2) # (k, m)

## [1] 0.0175231
```

Kontinuerliga Slumpvariabler

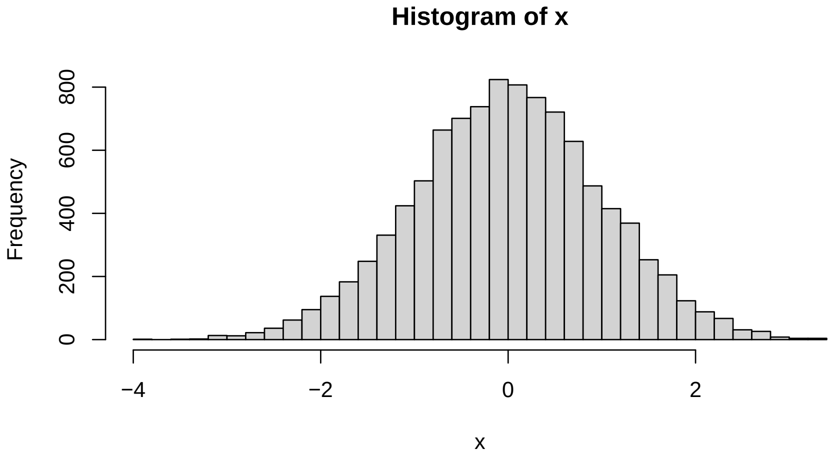
- Både binomialfördelning och Poissonfördelning kan bara anta heltalsvärden.
- En **kontinuerlig slumpvariabel** X kan anta decimalvärden, inte bara heltal.
  - Om vi har en kontinuerlig slumpvariabel är det meningslös att använda sannolikhetsfunktionen $P(X = x)$.
  - Vi har alltid $P(X = x) = 0$ oavsett värdet av x.

Histogram

```
hist(x)
```

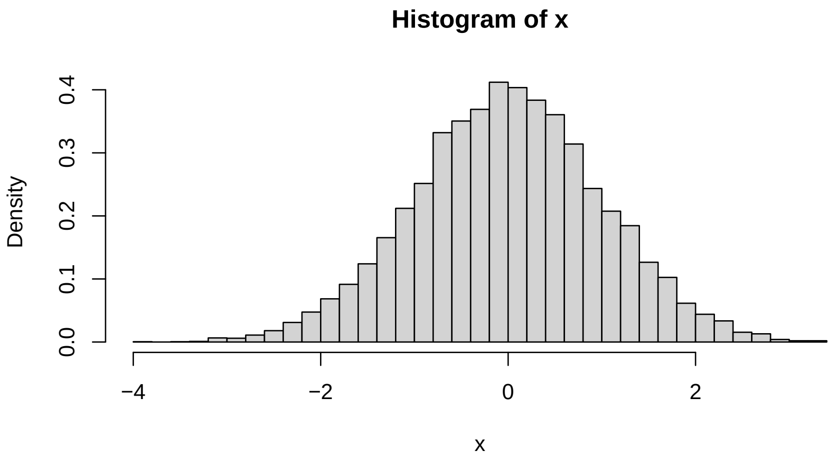


Histogram, $n = 10,000$



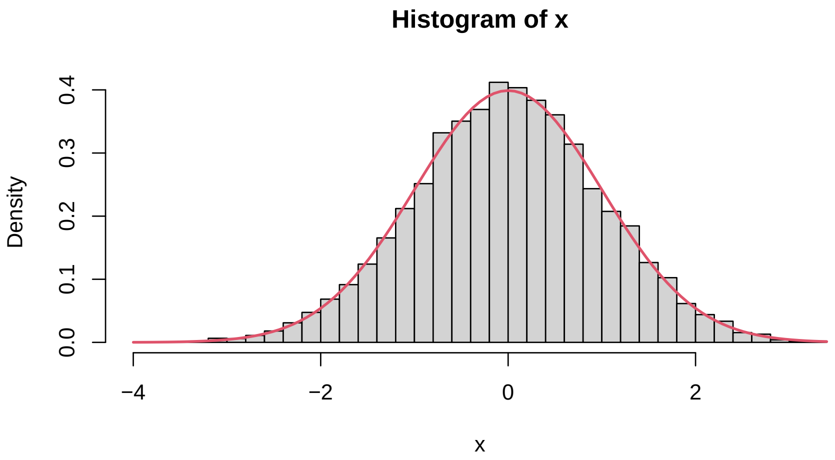
Histogram, $n = 10,000$

Vi skalar om histogrammet så att det får area 1.



Täthetsfunktion

Den kontinuerliga fördelningen kan beskrivas med en kontinuerlig **täthetsfunktion**:



Täthetsfunktion och Sannolikhet

För en **kontinuerlig slumpvariabel** med **täthetsfunktion** $f(x)$ beräknas sannolikheter med hjälp av integraler.

$$P(a < X \leq b) = \int_a^b f(x)\, dx.$$

Täthetsfunktionen kan aldrig vara negativ!

OBS!

1. Låt X vara en diskret slumpvariabel.

$$P(X \leq b) = P(X < b) + P(X = b) \neq P(X < b),$$

om $P(X = b) \neq 0$.

2. Låt X vara en kontinuerlig slumpvariabel.

$$P(X \leq b) = \int_a^b f(x)\,dx,$$

$$P(X < b) = \int_a^b f(x)\,dx = P(X \leq b).$$

Väntevärde och Standardavvikelse

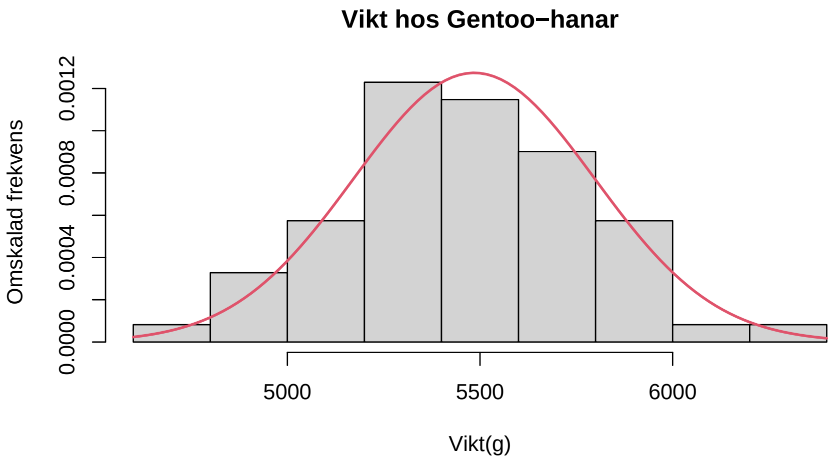
Väntevärde och varians för en kontinuerlig slumpvariabel beräknas med hjälp av integraler:

$$
\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x \cdot f(x)\, dx, \\
V(X) &= \int_{-\infty}^{\infty} [x - E(X)]^2 \cdot f(x)\, dx, \\
S(X) = \sqrt{V(X)} &= \sqrt{\int_{-\infty}^{\infty} [x - E(X)]^2 \cdot f(x)\, dx}.
\end{aligned}
$$

Ofta använder vi en känd fördelning för våra modeller, där vi inte behöver räkna ut väntevärde och varians för hand.

Våra Pingvinmätningar

Många fenomen i naturen har en symmetrisk fördelning:



Normalfördelning

Definition (den viktigaste fördelningen inom statistiken!)

Slumpvariabeln X är normalfördelad med parametrarna μ och σ, om dess täthetsfunktion är

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

då $-\infty < x < \infty$. Kodbeteckning: $X \sim N\left(\mu, \sigma^2\right)$.

Om X är normalfördelad med parametrarna μ och σ är

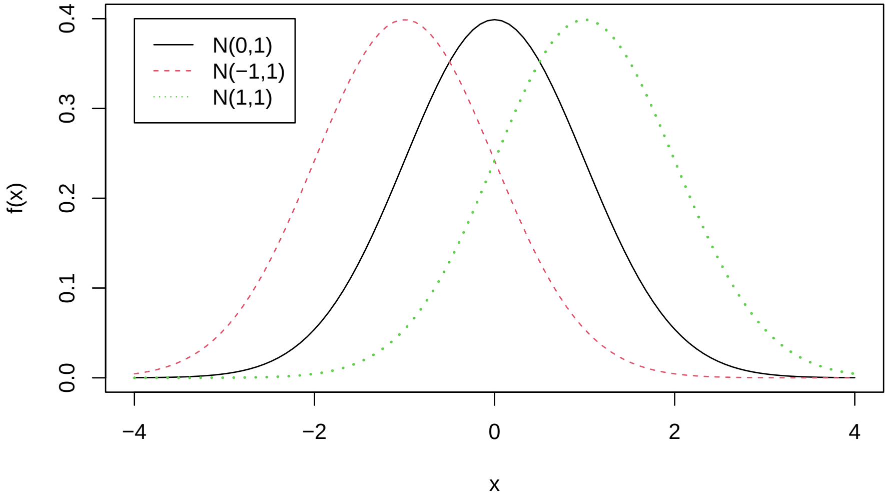
$$\begin{aligned} E(X) &= \mu, \\ V(X) &= \sigma^2, \end{aligned}$$

och

$$P(\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma) \approx 0.95.$$

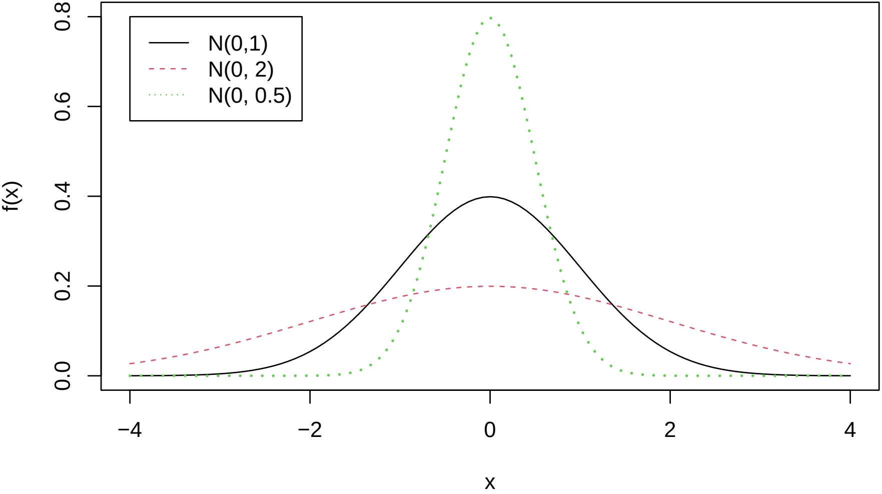
Täthetsfunktion av $X \sim N\left(\mu, \sigma^2\right)$

Effekten av μ: När väntevärdet μ ändras förskjuts täthetsfunktionen i sidled.



Täthetsfunktion av $X \sim N\left(\mu, \sigma^2\right)$

Effekten av σ: När standardavvikelsen σ ändras blir funktionen snävare eller bredare.



Sannolikheter

Fördelningsfunktionen

$$F(z) = P(X \leq z) \quad = \quad \int_{-\infty}^{z} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx$$

kan inte skrivas på sluten form! Sannolikheter för normalfördelade slumpvariabler räknas alltid ut med hjälp av en dator!

Om $X \sim N(0,1)$ är $P(X \leq 1.96)$

```
pnorm(1.96, 0, 1) # (z, mu, sigma)

## [1] 0.9750021
```

Pingvingmätning

Antag att vikt av en Gentoohane är normalfördelad med väntevärdet 5484.836 och standardavvikelsen 313.1586.
1. Vad är sannolikheten att en slumpmässigt vald Gentoohane väger mindre än 5000 g?

```
pnorm(5000, 5484.836, 313.1586) # (z, mu, sigma)

## [1] 0.06078559
```

2. Vad är sannolikheten att en slumpmässigt vald Gentoohane väger mer än 6000 g?

```
1 - pnorm(6000, 5484.836, 313.1586) # (z, mu, sigma)

## [1] 0.04997895
```

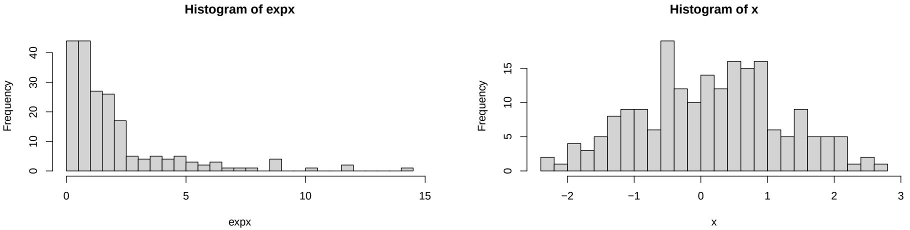
Pingvingmätning

Antag att vikt av en Gentoohane är normalfördelad med väntevärdet 5484.836 och standardavvikelsen 313.1586.
Vad är sannolikheten att en slumpmässigt vald Gentoohane väger mellan 5000 och 6000 g?

```
pnorm(6000, 5484.836, 313.1586) -
    pnorm(5000, 5484.836, 313.1586)

## [1] 0.8892355
```

Lognormalitet

Många mätvärden är inte normalfördelade men de se ut som normalfördelade efter logaritmering.



Exponentialfördelning

En kontinuerlig fördelning som används ofta för att beskriva tiden är **exponentialfördelningen**.

1. Tid från att en individ föds till att den dör
2. Tid mellan två hajattacker
3. Tid att vara en kund hos Spotify.

Definition

En slumpvariabel X är exponentialfördelad med parametern λ om dess täthetfunktion är

$$f(x) = \lambda e^{-\lambda x},$$

då $x > 0$. Kodbeteckning: $X \sim \text{Exp}(\lambda)$.

Om $X \sim \text{Exp}(\lambda)$,

$$E(X) = \frac{1}{\lambda}, \qquad V(X) = \frac{1}{\lambda^2}.$$

OBS!

Täthetfunktion av $X \sim \mathrm{Exp}\,(\lambda)$ är

$$f\,(x) \;=\; \lambda e^{-\lambda x}.$$

Ibland säger man att täthetfunktionen av $X \sim \mathrm{Exp}\,(\lambda)$ är

$$f\,(x) \;=\; \frac{1}{\lambda} e^{-x/\lambda}.$$

Då är

$$\begin{aligned} E\,(X) &= \lambda, \\ V\,(X) &= \lambda^2. \end{aligned}$$

Exponentialfördelning med R

The Exponential Distribution

Description

Density, distribution function, quantile function and random generation for the exponential distribution with rate `rate` (i.e., mean `1/rate`).

Usage

```
dexp(x, rate = 1, log = FALSE)
pexp(q, rate = 1, lower.tail = TRUE, log.p = FALSE)
qexp(p, rate = 1, lower.tail = TRUE, log.p = FALSE)
rexp(n, rate = 1)
```

Arguments

`x, q`	vector of quantiles.
`p`	vector of probabilities.
`n`	number of observations. If `length(n) > 1`, the length is taken to be the number required.
`rate`	vector of rates.
`log, log.p`	logical; if TRUE, probabilities p are given as log(p).
`lower.tail`	logical; if TRUE (default), probabilities are $P[X \leq x]$, otherwise, $P[X > x]$.

Details

If `rate` is not specified, it assumes the default value of `1`.

The exponential distribution with rate λ has density

$$f(x) = \lambda e^{-\lambda x}$$

for $x \geq 0$.

Exempel: Strömavbrott

Antag att tiden till nästa strömavbrott i månader är exponentialfördelad med parametern $\lambda = 0.25$. Beräkna sannolikheten för att nästa strömavbrott inträffar någon gång inom sex månader från nu.

$$P(X \leq 6) \quad = \quad \int_0^6 0.25 e^{-0.25x} dx \approx 0.78.$$

Med R:

```
pexp(6, 0.25) # (x, lambda), rate = lambda.

## [1] 0.7768698
```

Sammanfattning

1. Diskreta slumpvariabel (t.ex. binomial och Poisson).
   1. Sannolikheter, väntevärde, och varianser beräknas med summor.
2. Kontinuerlig slumpvariabel (t.ex. normal och exponential).
   1. Sannolikheter, väntevärde, och varianser beräknas med integraler.

Fördelning	Kodbeteckning	Väntevärde	Varians
Binomial	$\text{Bin}\left(n,p\right)$	np	$np\left(1-p\right)$
Poisson	$\text{Po}\left(m\right)$	m	m
Exponential	$\text{Exp}\left(\lambda\right)$	$1/\lambda$	$1/\lambda^2$
Normal	$N\left(\mu,\sigma^2\right)$	μ	σ^2