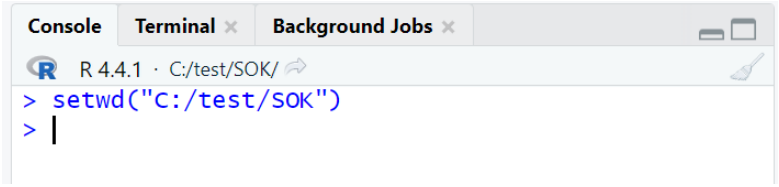
Datorlaboration 3, Statistisk översiktskurs (SÖK)

I den här labben går vi först igenom några kommandon för hur man kan arbeta med **normalfördelningen** i R, för att sedan kunna använda dessa kommandon för att ta fram konfidensintervall och utföra hypotestester. Materialet i labben relaterar till vad vi gjort på föreläsningarna 5-7. Vi kommer också säga något nedan om **t-fördelningen**, eftersom den fördelningen används när vi gör regression (F8-F10, labb 4). Vi gör också en plott med binomialfördelningen.

I appendix till labben (näst sista sidan) finns information om hur du får fram olika tecken på ditt tangentbord (bland annat tecknen ~ och |).

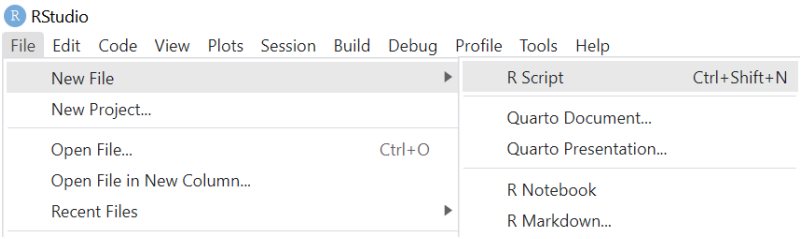
Innan själva uppgifterna börjar, gör följande (arbetskatalog, fil att spara i, ladda paket)

1. I Console i R, kör ditt **setwd**-kommando från labb 1 som sätter din arbetskatalog i R tlll din **SOK-mapp** på datorns hårddisk. Kommandot ser olika ut beroende på var på datorn din SOK-mapp finns.



Du kan alternativt göra detta steg genom menyerna i R, som vi gjorde i kapitel 6 i laboration 1, dvs: Välj menyn **Session** i R, sedan **Set Working Directory**, sedan **Choose Directory…** och klicka dig fram till mappen **SOK**.

2. Skapa en textfil där du kommer spara dina labb3-kommandon (ett ”R-script”) genom att i menyn i R välja **File** och sedan **New File** och sedan **R script**.



3. Labb 3 förutsätter att följande paket finns installerade: **openintro** och **mosaic**

Paket installeras via kommandot install.packages('packagename'), där packagename är paketnamnet. Om du inte redan gjort det, kör kommandot:

install.packages('openintro')

mosaic-paketet ska du redan ha installerat, om inte, installera det paketet också. Skriv sedan in följande library-kommandon i din textfil:

library(mosaic)

library(openintro)

Kör sen de två library-kommandona, ställ dig på första raden med ett library-kommado och tryck på Run, två gånger.

4. På samma sätt som i tidigare labbar är det bra att som första kommando i textfilen ha kommandot som bestämmer arbetskatalogen i R. Du kan kopiera in kommandot du körde i steg 1 ovan, in i din textfil (kopiera från Console eller från History-fliken i övre högra delen av RStudio.)

5. Spara din textfil genom att trycka på sparasymbolen  . Välj lämpligt namn. Spara din fil ofta.

1. Några användbara kommandon för fördelningar

Läs följande lista översiktligt, vi kommer främst använda **pnorm()** och **qnorm()**.

Fördelningar i R: r, p, d och q-funktionerna

R har en massa statistiska fördelningar inbyggda från start, och ytterligare många, många mer i olika R-paket. För varje fördelning finns det fyra typer av funktioner:

- **r**-funktionen som genererar slumptal från fördelningen, t ex `rnorm(n = 5, mean = 2, sd = 3)` genererar 5 slumptal från normalfördelningen $X \sim N(2, 3)$. Argumentet `sd` betyder standard deviation, dvs standardavvikelse.
- **p**-funktionen som beräknar sannolikheten att slumpvariabeln är mindre än ett visst tal, dvs $P(X \leq x)$ för något x. p-funktionen har fått sitt namn från engelskans **p**robability. `pnorm(q = 1, mean = 2, sd = 3)` beräknar sannolikheten att $X \sim N(2, 3)$ är mindre än 1.
- **d**-funktionen som för en diskret variabel X beräknar sannolikheten $P(X = x)$. För en kontinuerlig variabel beräknar d-funktionen täthetsfunktionen $f(x)$ i en given punkt x. Det är från det kontinuerliga fallet som d-funktionen fått sitt namn, d som i **d**ensity.
  - Kommandot `dnorm(x = 0, mean = 2, sd = 3)` beräknar täthetsfunktionens värde i punkten $x = 0$, dvs $f(0)$ för en $X \sim N(2, 3)$ variabel.
- **q**-funktionen beräknar p-kvantilen för en fördelning, dvs det värde x där $P(X \leq x) = p$. Vi kan t ex beräkna 25% eller 0.25-kvantilen för en normalfördelad variabel $X \sim N(2, 3)$ med kommandot `qnorm(p = 0.25, mean = 2, sd = 3)`.

På föreläsning har vi stött på beteckningen $\sim N(0, 1)$, nu har vi $\sim N(2, 3)$, vilket betyder att variabeln är normalfördelad (N), med medelvärde 2 och standardavvikelse 3.

På föreläsningarna och i kapitel 13 i kursboken har vi också sett användning av kommandona **pnorm()** och **qnorm()**. Med pnorm() kan vi gå från ett Z-värde (standardiserade normalfördelningen) till en sannolikhet, med qnorm() går vi från sannolikhet till Z-värde.

På nästa sida kommer två bilder från materialet på föreläsning 6, som också visar delar av den standardiserade normalfördelningstabellen (som också finns på Athena, bland extramaterial).

Notera att i pnorm- och qnorm-kommandona på nästa sida specificeras explicit att medel=0 (**mean=0**) och standardavvikelse=1 (**sd=1**), dvs att vi arbetar med den standardiserade normalfördelningen. Om vi inte anger dessa argument överhuvudtaget, kommer båda kommandona att anta just standardiserade normalfördelningen, men vi rekommenderar att ange värdena explicit. (Gällande **normTail**-kommandot ser vi att medelvärdesparametern heter **m** och standardavvikelseparametern **s**).

Ta fram sannolikhet under ett visst Z-värde

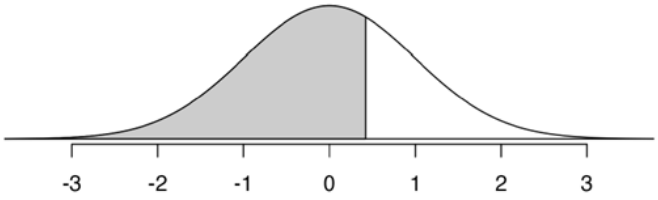
Med R:

Normal probabilities are most commonly found using statistical software which we will show here using R. We use the software to identify the percentile corresponding to any particular Z score. For instance, the percentile of $Z = 0.43$ is 0.6664, or the 66.64^{th} percentile. The `pnorm()` function is available in default R and will provide the percentile associated with any cutoff on a normal curve. The `normTail()` function is available in the **openintro** R package and will draw the associated normal distribution curve.

```
pnorm(0.43, mean = 0, sd = 1)
```

```
[1] 0.666
```

```
openintro::normTail(m = 0, s = 1, L = 0.43)
```



Med tabell:

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793

Ta fram det Z-värde som har en viss sannolikhet till vänster i fördelningen

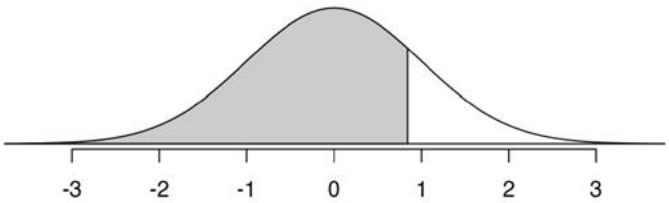
Med R:

We can also find the Z score associated with a percentile. For example, to identify Z for the 80^{th} percentile, we use `qnorm()` which identifies the **quantile** for a given percentage. The quantile represents the cutoff value. (To remember the function `qnorm()` as providing a cutozff, notice that both `qnorm()` and "cutoff" start with the sound "kuh". To remember the `pnorm()` function as providing a probability from a given cutoff, notice that both `pnorm()` and probability start with the sound "puh".) We determine the Z score for the 80^{th} percentile using `qnorm()`: 0.84.

```
qnorm(0.80, mean = 0, sd = 1)
```

```
[1] 0.842
```

```
openintro::normTail(m = 0, s = 1, L = 0.842)
```



Med tabell:

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327

Uppgift 1.1

Besvara följande frågor **med R och med hjälp av tabellerna ovan**: Hur stora andel av observationerna/sannolikheterna, för den standardiserade normalfördelningen ligger

a. under z=0.4
b. under z=0.89
c. mellan z=0.4 och z=0
d. mellan z= - 0.62 och z=0.62 (tips: även om du inte har tabellvärdet för negativa Z på föregående sida kan du använda att du vet att fördelningen är symmetrisk)

Uppgift 1.2

Besvara följande frågor **med R och med hjälp av tabellerna ovan**: Antag att individers kaloriintag av frukt och grönt är normalfördelat i befolkningen. Om en viss individ

e. äter mer frukt och grönt än 60% av andra individer, vilket Z-värde har individen (dvs: hur många standardavvikelser över medel finns individens konsumtion)?
f. äter mer frukt och grönt än 81% av andra individer, vilket Z-värde har individen (dvs: hur många standardavvikelser över medel finns individens konsumtion)?
g. äter mer frukt och grönt än 19% av andra individer, vilket Z-värde har individen (dvs: hur många standardavvikelser under medel finns individens konsumtion)? (för tabellsvaret: använd vad du vet om normalfördelningens form)

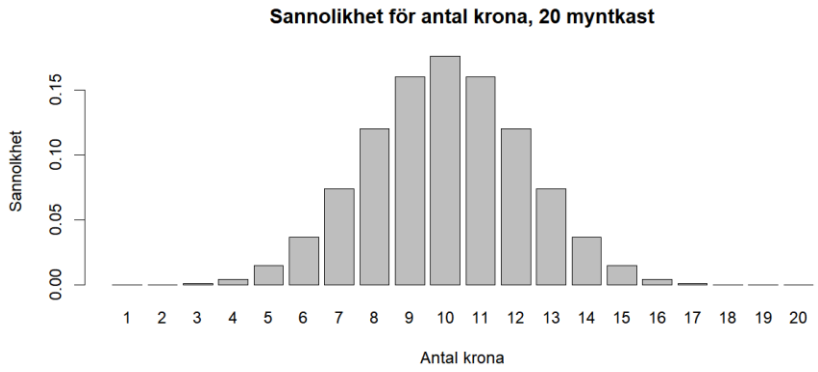
Uppgift 1.3 (vid intresse*)

Experimentera med normTail-kommandot för att rita olika varianter av normalfördelningen. Använd **?normTail** för att få upp en hjälp om kommandot. Om du exempelvis lägger till parametern U=.. kan du få upp två vertikala streck, och åskådliggöra exv. ett intervall. Om du har installerat openintro-paketet ska du kunna använda normTail-kommandot ”direkt” utan tillägget **openintro::** i början av kommandot.

Illustration (vid intresse*)
I kursen approximerar vi binomialfördelningen med normalfördelningen när vi arbetar med inferens gällande populationsandelar (F5-F7, Ö2, uppgifter nedan). Här ska vi enbart illustrera binomialfördelningen med en graf (som på F6, s. 14). Vi använder R-kommandot för densitetsfunktionen för binomialfördelningen, **dbinom()**. Vi tänker oss 20 myntkast där sannolikheten för krona är 50%. Vi tar fram den teoretiska fördelningen av antal krona (motsvarande om vi skulle upprepa 20 kast ett mycket stort antal gånger). Vi skapar en vektor med talen 1-20, sedan används dbinom() för att ta fram sannolikheten för varje värde av x, sedan används kommandot barplot (som finns med sist i labb 2), med olika argument, för att rita kurvan:

```
xvalues=seq(1:20)
yvalues = dbinom(xvalues,size=20, p=0.5)
```

barplot(yvalues, xlab = "Antal krona", ylab="Sannolkhet", main="Sannolikhet för antal krona, 20 myntkast", names.arg=xvalues)



2. Konfidensintervall för en andel

Som exempel på en andelsvariabel (eng: proportion) tänker vi oss andelen vuxna svenskar (18+) som sopsorterar. Vi är intresserade av hur stor andelen är i denna befolkningsgrupp. Vi gör en undersökning med 600 slumpvis utvalda individer och eftersom vi valt från hela den vuxna befolkningen kan vi anta att observationerna är oberoende av varandra.

60% av de tillfrågade säger att de sopsorterar, övriga att de inte sopsorterar, vi antar att svaren är sanningsenliga (olika typer av problem med undersökningar diskuteras på föreläsning 11).

Vi har alltså urvalsstorleken n=600, och **punktskattningen** $\hat{p}$=0.6

- För **andelar** ska vi verifiera **success-failure condition** (kap 16.2.1):
  - Stickprovsfördelningen för $\hat{p}$, från ett urval av storlek n från en population med sann andel p, är ungefärligen normalfördelad när vi förväntar oss minst 10 lyckade och 10 misslyckade utfall:
  - $np \geq 10$ och $n(1-p) \geq 10$ (i praktiken: använd $\hat{p}$)

Uppgift 2.1. Verifiera success-failure condition

Vi behöver beräkna standardfelet (standard error, föreläsning 6) i våra data, vilket används som ett mått på variationen i skattningen av populationsandelen.

- För en **andel** $\hat{p}$ kan man härleda att standardfelet är

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- där $\hat{p}$ är den skattade andelen

I R kan vi använda:

```
> se <- sqrt(0.6*(1-0.6)/600)
>
```

Vi behöver välja konfidensnivå, vi väljer 90% (vilket ger α=0.05, ”fem procent av sannolikheten i vardera svansen”) – se föreläsning 6 (s. 27) - då kan vi ta fram kritiskt Z-värde med pnorm() eller med standardnormalfördelningstabellen.

```
> qnorm(0.05, mean=0, sd=1)
[1] -1.644854
>
```

5% av sannolikheten ligger till vänster om z= - 1.644854, och 5% till höger om z = 1.644854 (normalfördelningen är symmetrisk).

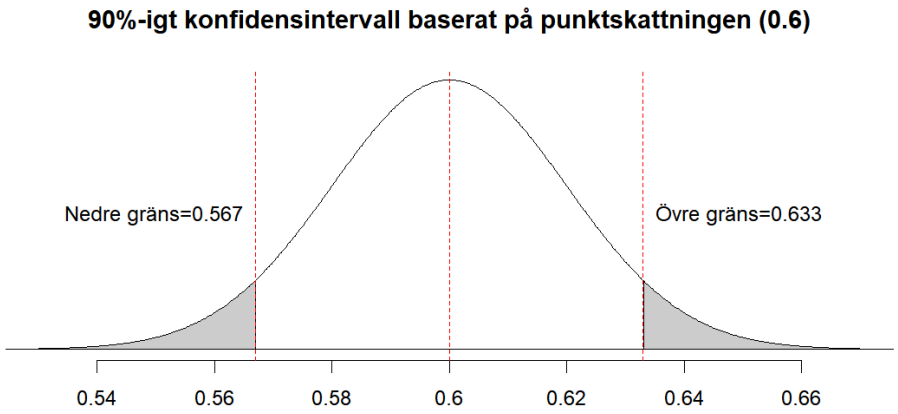
Vårt 90%-iga konfidensintervall ges mao av:

$\hat{p} \pm 1.644854 \times se = 0.6 \pm 1.644854 \times 0.02$, som ger intervallet (avrundat till tre decimaler) (0.567, 0.633)

```
> 0.6-0.02*1.644854
[1] 0.5671029
> 0.6+0.02*1.644854
[1] 0.6328971
>
```

Vi ritar en graf. Experimentera gärna med liknande R-kod, där första kodraden ritar upp normalfördelningen, gråmarkerar de två områdena och sätter en rubrik, andra raden ritar tre vertikala röda streck och de två sista raderna skriver ut text i grafen.

```
> normTail(m=0.6, s=0.02, L=0.567, U=0.633, main="90%-igt konfidensintervall baserat på punktskattningen (0.6)")
> abline(v = c(0.567, 0.6, 0.633), col = "red", lty = 2)
> text(0.633, 10, "Övre gräns=0.633", pos = 4)
> text(0.567, 10, "Nedre gräns=0.567", pos = 2)
>
```



Ovan användes ett objektnamn ”se” för standardfelet. Det fungerar bra men kan bli förvillande om du gör flera liknande uppgifter. Om standardfelet i nästa uppgift också döps till se skrivs det tidigare värdet över, vilket kan ge felaktigheter om du skulle vilja gå tillbaka till tidigare uppgift. Kanske är ett något längre namn det mest lämpliga, så att olika uppgifter inte blandas ihop.

Uppgift 2.2

Följ metoden ovan och ta fram ett 95%-igt konfidensintervall, mha. R, för väljarandelen som stöder Kristersson (från F6):

- En Novusundersökning omfattade 509 personer
- "Vilket förtroende har du för statsministern?"
- 204 indidivider svarade "ganska stort" eller "mycket stort"
- Beräkna ett 95%-igt konfidensintervall för andelen bland alla väljare som har ganska stort eller mycket stort förtroende för Kristersson

Kontrollera också success-failure condition. Tolka konfidensintervallet. Jämför med svaret på föreläsning 6.

3. Hypotestest för en andel

Din vän säger att **mer än** 65% av en viss grupp (population) är emot införande av Euron i Sverige. Du är skeptisk och hävdar att det är 65% mot. Du drar ett slumpvis urval av 336 individer ur gruppen, och vi kan anta att dessa kan hanteras som oberoende observationer. I ditt urval svarar 70% ”mot Euro”, och 30% svarar ”för Euro”. Sätt upp ett hypotestest och utvärdera hypoteserna med signifikansnivån α=0.05. Tolka resultatet.

Med bas i föreläsning 7 kan vi formulera följande:

Nollhypotes: Populationsandelen (p) som är emot Euroinförande är 0.65 (ditt skeptiska perspektiv).

Alternativ: Populationsandelen (p) som är emot Euroinförande är större än 0.65 (din kompis hypotes).

Vi kallar värdet i nollhypotesen p_0, mao, p_0= 0.65. Hypoteserna kan också uttryckas som:

$$H_0 : p = p_0$$

$$H_A : p > p_0$$

Uppgift 3.1. Verifiera success-failure condition

Uppgift 3.2 Ta fram n, punktskattningen $\hat{p}$ från vårt urval, gör standardfelsberäkningen i R och verifiera att standardfelet blir 0.025.

Vi har ett ensidigt hypotestest för en andel (F7, sid 27) och ska

- **Ta fram värdet på testvariabeln:**

$$Z_{\text{obs}} = \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

Uppgift 3.3. Verifiera att testvariabelns värde, Z_{obs}, blir 2

Med vår testvariabel standardiserad på detta sätt kan vi jämföra vårt observerade Z-värde (Z_{obs}) med det **kritiska värde** vi får från den standardiserade normalfördelningen, detta värde beror i sin tur på vår valda signifikansnivå.

- **Signifikansnivå** α**:** sannolikheten att ta fel beslut, dvs att förkasta nollhypotesen fast vi inte borde (direkt kopplat till kritiska värdet)

Ovan har vi valt α=0.05.

Uppgift 3.4. Med qnorm()-kommandot, ta fram det Z-värde för vilket 95% av sannolikheterna/ observationerna återfinns för lägre Z-vörden (”ligger till vänster i fördelningen”) (du ska få, om vi avrundar till tre decimaler, 1.645)

Mer om ensidigt test:

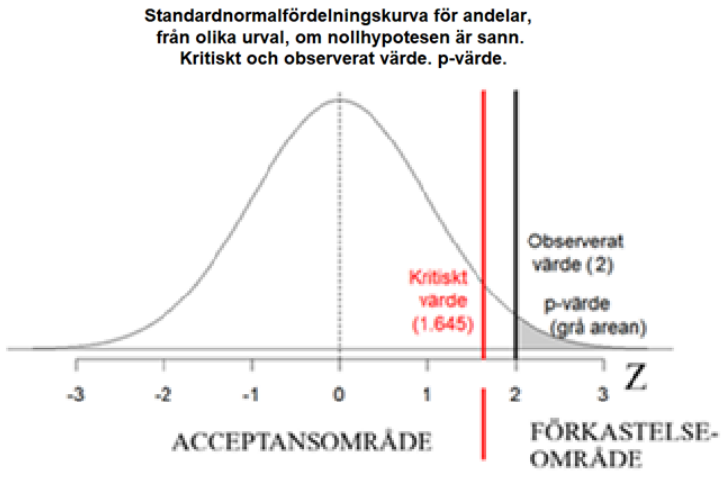
Kompisens påstående gäller att Euromotståndet är **större än** ett visst värde (en viss populationsandel), vi ska därför bara ta hänsyn till sannolikheten i högra svansen av standardnormalfördelningskurvan, när vi utvärderar påståendet. **Om nollhypotesen är sann** – dvs andelen är 0.65 – är sannolikheten att observera ett Z-värde över 1.645 fem procent. Det vill säga, sannolikheten för ett felbeslut, med detta värde som kritisk gräns, skulle vara fem procent. Vi accepterar denna risk. Om det observerade Z-värdet är 1.645 eller högre förkastar vi därför nollhypotesen (att andelen i populationen skulle vara 0.65) och tar istället det höga Z-värdet som evidens för alternativhypotesen.

Vi tar fram en graf, med följande kommandon (liknar uppgift 2.1 ovan):

(Text har också lagts in i grafen ”manuellt”).

```
normTail(m=0, s=1, U=(0.7-0.65)/0.025, main="Standardormalfördelningskurva för andelar, \n från olika urval, om
nollhypotesen är sann. \n Kritiskt och observerat värde. p-värde.", cex.main=1)
abline(v = 0, col = "black", lty = 2)
abline(v = 1.645, col = "red", lty = 1, lwd=3)
abline(v = 2, col = "black", lty = 1, lwd=3)
text(1.64, 0.07, "Kritiskt \n värde \n (=1.645)", pos = 2, col="red")
text(2, 0.15, "Observerat \n värde (2)", pos = 4, col="black")
text(2.2, 0.05, "p-värde \n (grå arean)", pos = 4, col="black")
```

Några saker att notera i dessa kommandon är att radbrytning lagts in i text som visas, med **\n** i själva texten. **cex.main()** i rubriken bestämmer storlek på rubriken. U-värdet i första kommandot har också standardiserats (dra ifrån p_0, dividera med standardfel). abline-kommandona ritar vertikala streck för noll, kritiskt Z-värde och observerat Z-värde.



Uppgift 3.5 Ska vi förkasta nollhypotesen? (resonera med hjälp av texten ”Mer om ensidigt test” ovan)

Uppgift 3.6 Använd lämpligt R-kommando för att ta fram p-värdet (grå arean), dvs

- **p-värde:** Sannolikheten att observera ett värde som är minst lika extremt som det observerade, givet att nollhypotesen är sann

Du ska få värdet 0.02275.

4. Kommandot prop.test(), skillnad mellan två andelar etc. – läs vid intresse

prop.test() automatiserar en del av de uppgifter vi gått igenom ovan.

Testa följande

```
> prop.test(x=360, n=600, conf.level = 0.9)

        1-sample proportions test with continuity
        correction

data:  360 out of 600
X-squared = 23.602, df = 1, p-value = 1.185e-06
alternative hypothesis: true p is not equal to 0.5
90 percent confidence interval:
 0.5658840 0.6331955
sample estimates:
  p
0.6
```

I detta fall tar kommandot (bland annat) fram ett konfidensintervall baserat på en skattning av en populationsandel och baserat på urvalsstorlek 600 och med 360 ”ja-svar” – samma frågeställning som vi hade med sopsorteringen ovan (0.6 i observerad andel, dvs 360 av 600). Konfidensintervallet blir marginellt annorlunda eftersom metoden vi använde ovan och metoden i prop.test skiljer sig (något) åt.

I kursen är det viktigt att förstå de olika stegen i beräkningar av konfidensintervall och i hypotestest – av denna anledning använde vi inte prop.test() ovan.

Skatta skillnad mellan två populationsandelar (F7, s. 28, inte tentamaterial men bra att veta)

Antag att du gjort din undersökning av sopsortering vid två tillfällen (exv. 2020 och 2025), med slumpvisa urval från samma population och oberoende observationer. Du intervjuade 600 personer båda gångerna. 2020 svarade 300 personer ”Ja”, 2025 360 personer. Skillnaden i skattad andel är alltså 0.1 (från 0.5 till 0.6). Följande kommando tar (bland annat) fram ett konfidensintervall för skillnaden:

```
> prop.test(x=c(360, 300), n=c(600,600), conf.level = 0.9)

        2-sample test for equality of proportions with
        continuity correction

data:  c out of c360 out of 600300 out of 600
X-squared = 11.721, df = 1, p-value = 0.0006181
alternative hypothesis: two.sided
90 percent confidence interval:
 0.05132773 0.14867227
sample estimates:
prop 1 prop 2
   0.6    0.5
```

Om vi jämför med kodrutan ovan ser vi att konfidensintervallet för skillnaden är bredare. Vi har mer osäkerhet. Variationen i skattningen av en skillnad är högre än variationen i skattningen av en enda andel.

Samma slutsats – mer variation i skattningen – vilket är relevant för inferens - gäller också för skattningar av skillnader mellan numeriska variabler i allmänhet.

5. Konfidensintervall och hypotestest numeriska variabler

Gå igenom exemplet i slutet av föreläsning 7 (s. 24-26).

Om vi först bortser från hypotestest och vill ta fram ett konfidensintervall, har vi de relevanta värdena från vårt urval på sid 24.

Vi kan ta fram (exv.) ett 95%-igt konfidensintervall med följande formel, från F6:

- 95%-igt konfidensintervall $(\alpha/2 = 0.025)$ ges av:

$$\bar{x} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}} = \bar{x} \pm Z_{0.025} \frac{s}{\sqrt{n}} = \bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$$

där 1.96 kommer från pnorm()-kommandot / standardnormalfördelningsrabellen.

5.1 (Träning inför tenta)

Räkna ut konfidensintervallet manuellt med formeln ovan och verifiera ditt svar med motsvarande R-beräkning. Du bör få (3.608, 4.392). Tolka intervallet.

5.2 (Relevant för inlämningsuppgiften, del 3, och som träning inför tenta)

Gå igenom F7, speciellt från sid 14, och exemplet (s. 24-26).

Ställ upp noll- och alternativhypotes.

Från urvalsundersökningen, ta fram punktskattning $\bar{x}$, standardavvikelse s, standardfel, och observerat Z-värde, på samma sätt som på F7, s.24, där μ_0 är populationsmedelvärdet som antas i nollhypotesen.

Träna på att rita en bild (papper och penna), motsvarande bilden på sid 26.

Baserat på valt α (vald signifikansnivå för hypotestestet), ta fram kritiskt Z-värde.

Utvärdera hypotesen (förkastas nollhypotesen?)

Träna på att formulera i ord.

6. Normalfördelningen och t-fördelningen – bra att veta (gäller kontinuerliga numeriska variabler i allmänhet, inte andelar)

Följande bild från boken kap 19 visar att **t-fördelningen** (som nämndes på F6, sista sidan), för små urval, har bredare svansar än normalfördelningen.

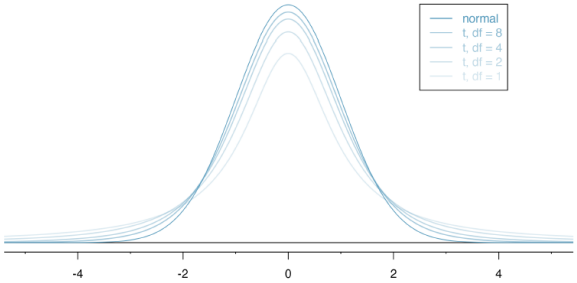


Figure 19.9: The larger the degrees of freedom, the more closely the t-distribution resembles the standard normal distribution.

Degrees of freedom: df.

The degrees of freedom describes the shape of the t-distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal distribution.

When modeling $\bar{x}$ using the t-distribution, use $df = n - 1$.

Antag att vi har n=8 observationer (slumpvis utvalda från populationen, oberoende).

Till vänster om Z=-2 i standardnormalfördelningen finns

```
> pnorm(-2,m=0,s=1)
[1] 0.02275013
```

dvs 2.27% av sannolikheten/observationerna.

Till vänster om samma punkt, i t-fördelningen med n-1=7 frihetsgrader, finns

```
> pt(-2, df=7)
[1] 0.04280966
>
```

dvs 4.28% av sannolikheten/observationerna.

- En lärdom är att om vi, för små n, skulle använda normalfördelningen, när vi borde använda t-fördelningen, kan vi dra fel slutsatser om konfidensintervall och om hypoteser. Exempelvis är det lättare att (felaktigt) förkasta en nollhypotes med normalfördelningen än med t-fördelningen. För n>30 blir det i praktiken ingen skillnad.

- Vi såg R-funktionen **pt()**, som fyller samma funktion som pnorm(), men för t-fördelningen.

- Att förklara frihetsgradsbegreppet ingår inte i kursen.

-

- Inom regressionsanalys används t-fördelningen när vi gör hypotestest om regressionskoefficienter och läser regressionstabeller (F8-F10).

Kortkommandon etc.

Description of Shortcuts and Special Characters in R

Description	Windows/Linux	Mac
Run code	Ctrl + Enter	⌘ + Enter
Run code and stay on current line	Alt + Enter	Option + Enter
Select All	Ctrl + A	⌘ + A
Copy	Ctrl + C	⌘ + C
Paste	Ctrl + V	⌘ + V
Cut selected text	Ctrl + X	⌘ + X
Delete line	Ctrl + D	⌘ + D
Undo	Ctrl + Z	⌘ + Z
Redo	Ctrl + Shift + Z	⌘ + Shift + Z
Select multiple lines/area	Shift + arrow key	Shift + arrow key
Duplicate line/area	Ctrl + Shift + D	⌘ + Shift + D
Scroll up/down	Ctrl + up/down arrow key	⌘ + up/down arrow key
Go to the top	Ctrl + Home	⌘ + Home
Go to the bottom	Ctrl + End	⌘ + End
Go to line	Shift + Alt + G	⌘ + Shift + Option + G
Save	Ctrl + S	⌘ + S
Save all documents	Ctrl + Alt + S	⌘ + Option + S
Open document	Ctrl + O	⌘ + O
$ symbol	Ctrl + Alt + 4	Option + 4
Vertical bar/pipe (logical OR): |	Ctrl + Alt + < or Alt gr + <	Option + 7
Assignment (<-)	Alt + -	Option + -
Zoom out/in	Ctrl + -/+	⌘ + -/+
Interrupt running code	Esc	Esc
Clear Console	Ctrl + L	⌘ + Option + L
Left square bracket [	Ctrl + Alt + 8 or Alt gr + 8	Option + [or Option + 8
Right square bracket]	Ctrl + Alt + 9 or Alt gr + 9	Option +] or Option + 9
Left curly brace {	Ctrl + Alt + 7 or Alt gr + 7	Option + Shift + 8
Right curly brace }	Ctrl + Alt + 0 or Alt gr + 0	Option + Shift + 9
Slash /	Shift + 7	Shift + 7
Backslash \	Alt gr + + or Ctrl + Alt + +	Option + Shift + 7
Tilde symbol ~	Alt gr + ~(key near Å)	Option + ~(key near Å)
Search expression	Ctrl + F	⌘ + F
Arrange indentation	Ctrl + I	⌘ + I
Show list of common shortcuts	Shift + Alt + K	Option + Shift + K

Denna version av dokumentet: 260407

Materialet i Statistisk översiktskurs har tagits fram av Ulf Högnäs och Anders Fredriksson, med inspiration och ibland direkt användande av material från andra kurser och personer, bland annat kurserna Statistik och dataanalys 1-3, med material av Michael Carlson, Ellinor Fackle Fornius, Jessica Franzén, Oskar Gustafsson, Oscar Oelrich, Mona Sfaxi, Karl Sigfrid, Mattias Villani, Valentin Zulj, med flera.