

Tölfræði og gagnavinnsla (STÆ209G), lokapróf

4. maí, 2018

Nafn: _____ HÍ notendanafn: _____

Á prófinu eru fjögur skrifleg dæmi og 20 krossasurningar. Fyrir hverja rétta krossasurningu fást 2.5 stig. Tvö skriflegu dæmanna eru 15 stig og tvö þeirra eru 10 stig. Alls er prófið 100 stig.

Skrifið lausnina ykkar á skriflegu dæmunum á prófið. Ef þið þurfið meira pláss skulið þið biðja um auð blöð hjá prófverði (þið getið líka beðið um auð blöð sem þið getið notað sem krassblöð). Ef hluti af lausninni ykkar er á auka blöðum þurfið þið að tilgreina það við dæmið á prófinu og merkja auka blöðin vel og vandlega.

Aðeins er eitt rétt svar við hverja krossasurningu. Ekki verður dregið frá fyrir röng svör. Á blaðsíðu 17 er krossasurningasvarblað. Merkið við þau svör sem þið teljið vera þau réttu á krossasurningasvarblaðinu. Blaðsíður 9 - 16 verða ekki skoðaðar við yfirferð prófsins og því er mjög mikilvægt að þið fyllið út töfluna á blaðsíðu 17.

Leyfilegt er að vera með allar gerðir vasareikna í prófinu en önnur hjálpargögn eru ekki leyfileg. Formúlublað og töflur fylgja prófinu. Vinsamlegast skrifið ekki á formúlublaðið og töflurnar svo hægt verði að nota þau aftur.

Gangi ykkur vel:)

Skrifleg dæmi

Dæmi 1 (15 %)

Neytendasamtökin ákváðu að standa fyrir rannsókn þar sem þrjár mismunandi tegundir af lyftidufti (A, B og C) voru bornar saman í þeim tilgangi að athuga hvort munur væri á virkni þeirra. Rannsóknin fór þannig fram að sex kökur voru bakaðar með hverri gerð af lyftidufti. Rúmmál kakanna var mælt og eftirfarandi stærðir voru einnig reiknaðar út

$$SS_T = 2039.72, \quad MS_E = 82.40$$

a) (6 %)

Fyllið út eftirfarandi fervikasummutöflu (þið eigið að skrifa sex tölur inn í töfluna):

Fervikasummur	Frígráður	Meðalfervikasummur
$SS_{Tr} = 803.72$	2	$MS_{Tr} = 53.58$
$SS_E = 1236$	15	$MS_E = 82.40$
$SS_T = 2039.72$	17	

b) (9 %)

Kannið með viðeigandi tilgátuprófi hvort munur sé á meðalrúmmáli kaka eftir því hvaða lyftiduftstegund er notuð. Sýnið öll skrefin 6. Notið $\alpha = 0.05$. Ef þið gátuð ekki fyllt út MS_{Tr} í töflunni hér að ofan skulið þið nota $MS_{Tr} = 500$.

Meðaltal próf tveggja eða fleirri þýða

Hæstu ásættanlegu villulíkur = 0.05 = alpha

H_0 er að ekki sé munur á meðalrúmmáli kaka eftir lyftiduftstegund $F = MSTr / MSTe = 0.65$

$F > F(1-\alpha, (a-1, N-a)) = 0.59$ þannig við höfnum H_0

Það er munur á rúmmáli eftir því hvaða lyftiduft er notað

Lausn, áfrh.:

Dæmi 2 (15 %)

Baldur er mikill áhugamaður um laun. Hann safnar gögnum um laun og hversu langa skólagöngu fólkið átti að baki frá 100 launþegum í stóru fyrirtæki með það að markmiði að búa til líkan sem getur spáð fyrir launum með menntunarstigi (mælt í fjölda ára í skóla að loknu grunnskólaprófi). Launþegarnir sem hafa lægst menntunarstig í gögnunum hans Baldurs fóru strax út á vinnumarkaðinn eftir grunnskólapróf en þeir sem hafa hæst menntunarstig eiga 9 ára skólagöngu að baki eftir grunnskólapróf. Baldur byrjaði á því að teikna upp gögnin og sá að sambandið er línulegt. Hann reiknaði auk þess út nokkrar lýsistærðir sem sjá má hér að neðan:

Breyta	Meðaltal	Staðalfrávik
Laun	660020	150100
Menntun (ár frá grunnskólaprófi)	5.5	2.2

Hann reiknaði einnig út fylgnistuðulinn milli launa og menntunarstigs (mælt í fjölda ára í skóla frá grunnskólaprófi) og fékk út $r = 0.55$.

a) 5%

Finnið jöfnu aðhvarfslínu minnstu kvaðrata fyrir líkanið hans Baldurs.

$$b_1 = r \cdot s_y / s_x = r \cdot \text{staðallaun} / \text{staðalmenntun} = 37525$$

$$b_0 = m_y - b_1 \cdot m_x = 660020 - 37525 \cdot 5.5 = 453632.5$$

$$y = 453632.5 + 37525x + \text{eps}$$

b) 2%

Hvert er skýringarhlutfall líkansins og hvað segir það okkur?

$$\text{skýringarhlutfall} = r^2 = 0.302$$

umþb 30% af breytileika launa skýrist af menntun

c) 2%

Hversu mikið hækka laun launþega fyrirtækisins fyrir hvert ár í skóla umfram grunnskólapróf samkvæmt líkaninu?

sjá b1, 37.525kr á ár

d) 3%

Georg nokkur hefur lokið 5 háskólagráðum og eyddi heilum 17 árum í skóla eftir grunnskólapróf. Notið líkanið til að spá fyrir um laun Georgs. Er spáin áreiðanleg? Rökstyðjið.

$$17 \cdot 37.525 = 637925$$

spáin er óáreiðanleg þar sem tímabilið er utan mældra gagna

e) 3%

Gunnar vill kanna með viðeigandi tilgátuðrófi hvort fylgnistuðullinn sé frábrugðinn núlli. Hvert er gildið á prófstærðinni sem Gunnar á að nota?

$$T = (r \cdot \sqrt{n-2}) / \sqrt{1-r^2}$$

$$0.55 \cdot 9.89 / 0.835 = 6.51$$

Dæmi 3 (10 %)

Gera má ráð fyrir að innhringingar í bilanabjónustu ákveðins fyrirtækis hér í bæ séu að meðaltali 3.5 á mínútu.

Þið megið svara b), c) og d) lið með R-skipunum.

a) 2%

Hvaða líkindadreifingu er eðlilegt að nota til að lýsa fjölda innhringinga á mínútu í bilanabjónustu fyrirtækisins og hvert er gildið á stika hennar?

poisson dreifing

b) 2%

Hverjar eru líkurnar á að á næstu mínútu hringi nákvæmlega einn viðskiptavinur í bilanabjónustuna?

dpois(1, 3.5)

c) 3%

Hverjar eru líkurnar á að á næstu mínútu hringi fleiri en einn viðskiptavinur í bílanaþjónustuna?

$$1 - \text{ppois}(1, 3.5)$$

d) 3%

Hverjar eru líkurnar á að á næstu tveimur mínútum hringi nákvæmlega fimm viðskiptavinir í bílanaþjónustuna?

$$\text{dpois}(5, 7)$$

Dæmi 4 (10 %)

Tóbaksframleiðandinn "Karólína Tóbak" auglýsti að mest seldu sígarettarnar þeirra innihéldu í mesta lagi 40mg af nikótíni. Talsmaður neytenda gerði prufu á 20 sígarettum völdum af handahófi úr mismunandi pökkum og fékk út að meðalnikótín magn í sígarettunum tuttugu var 42mg og var staðalfrávik 4mg. Gera má ráð fyrir að nikótíninnihald fylgi normaldreifingu. Framkvæmið viðeigandi tilgátupróf til að kanna hvort meðalnikótíninnihald sé frábrugðið 40mg. Sýnið öll skrefin sex. Notið $\alpha = 0.05$.

Tilgátupróf fyrir meðaltal yfir stakt þýði

$0.05 = \alpha$

H_0 er að meðaltal = 40

$Z = 2 - (0.835 * (4 / 4.47)) = 1.252$

Erum á höfnunarsvæði

Við höfnum því að meðaltalið sé frábrugðið 40mg

Krossaspurningar

Lalli langstökkvari er kennari í fjölmönnum skóla hér í borg. Hann hefur mikinn áhuga á langstökki og ákvað hann að rannsaka langstöcksgetu nemenda skólans. Hann hélt langstöcksmót þar sem allir nemendur skólans tóku þátt og mældi hann lengd stökka nemendanna. Gerum nú ráð fyrir að lengd stökka meðal nemenda skólans fylgi normaldreifingu með meðaltal 4 metrar og staðalfrávik 1 metri.

Spurningar 1-5 fjalla um langstöcksmótið í skólanum hans Lalla.

Spurning 1

Af hvaða gerð er breytan sem Lalli mælir?

- a) Samfelld flokkabreyta.
- b) Strjál talnabreyta.
- c) Strjál flokkabreyta.
- ☒ d) Samfelld talnabreyta.
- e) Röðuð flokkabreyta.

Spurning 2

Hverjar eru líkurnar á að nemendi, valinn af handahófi, stökkvi lengra en 4.5 metra?

- a) 0%
- b) um 6%
- c) um 31 %
- ☒ d) um 69 %
- e) um 94.3%

$$pnorm(4.5, 4, 1)$$

Spurning 3

Íris, nemandi í skólanum, er mikil keppnismanneskja og stefnir hún á að eiga eitt af 1% lengstu stökkunum. Hversu langt þarf Íris að stökkva að lágmarki til að ná því markmiði?

- a) um 1.67 metra
- b) um 5.28 metra
- c) um 5.33 metra
- ☒ d) um 6.33 metra
- e) um 7.11 metra

$$qnorm(0.99, 4, 1)$$

Spurning 4

Í skólanum hans Lalla eru 10 nemendur í bekk. Sé bekkur valinn af handahófi, hverjar eru líkurnar á að meðallengd stökkva nemenda bekkjarins sé undir 4.5 metrum?

- a) 0%
- b) um 6%
- c) um 31 %
- d) um 69 %
- ☒ e) um 94.3%

$$m = N(4, 0.1)$$

Spurning 5

Lalli valdi 100 þátttakendur úr langstökkskeppninni af handahófi og spurði þá hvort þeir hafi gaman að langstökki. 62 aðspurðra sögðust hafa gaman að langstökki. Lalli notaði gögnin til að smíða 95% öryggisbil fyrir hlutfall nemenda sem hafa gaman að langstökki. Hvert er **neðra** mark öryggisbilsins?

- a) 0.45
- b) 0.52
- ☒ c) 0.58
- d) 0.61
- e) 0.71

$$0.62 - 0.835 * \text{sqrt}((0.62*0.38)/100)$$

Spurning 6

Linda lífefnafræðingur er að rannsaka ákveðið efnahvarf. Hún mældi 50 sinnum tímann sem það tekur fyrir efnahvarfið að ljúka. Linda teiknaði stuðlarit af tímanum og sá hún að dreifingin er vinstri skekkt einkryppudreifing. Hver af eftirfarandi fullyrðingum er rétt?

- ☒ a) Meðaltíminn er lægri en miðgildi tímans.
- b) Meðaltíminn er hærri en miðgildi tímans.
- c) Dreifni er góður mælikvarði á breytileika tímans.
- d) Meðaltal og miðgildi tímans eru jöfn.
- e) Dreifni er góður mælikvarði á miðju tímans.

Í þorpi í landi nokkru langt langt í burtu eru 9% þorpbúa með ákveðinn sjúkdóm sem ekki er smitandi.

Spurningar 7 - 9 fjalla um þennan sjúkdóm.

Spurning 7

Séu 4 þorpsbúar valdir af handahófi, hverjar eru líkurnar á að einn eða fleiri séu með sjúkdóminn?

- a) um 4%.
- b) um 19%.
- c) um 27%.
- ☒ d) um 31%.
- e) um 69%.

Spurning 8

Séu 20 þorpsbúar valdir af handahófi, hver af R-skipununum hér að neðan gefa líkurnar á því að fleiri en fimm þeirra séu með sjúkdóminn?

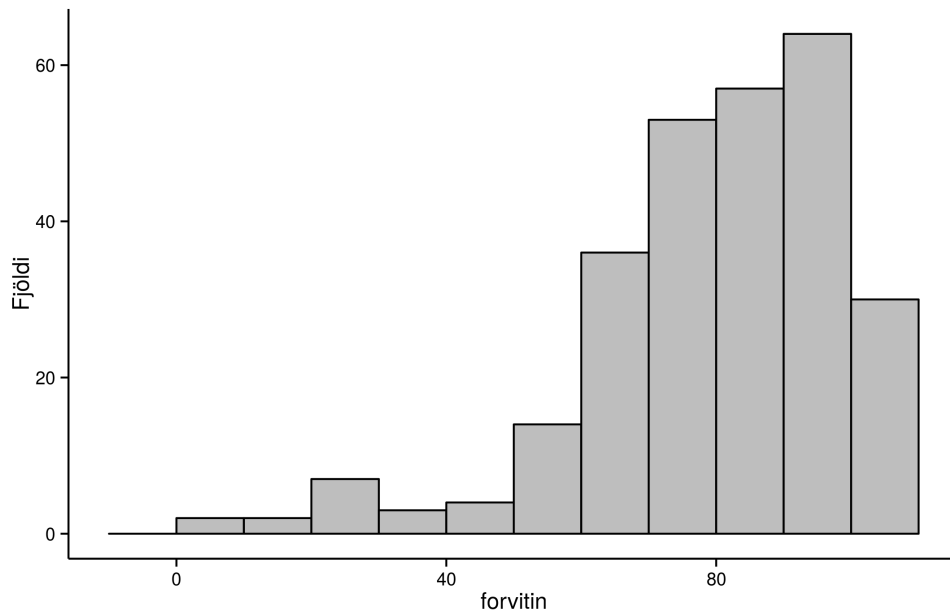
- ☒ a) `1-pbinom(5,20,0.09)`
- b) `pbinom(5,20,0.09)`
- c) `1-dbinom(4,20,0.09)`
- d) `dbinom(5,20,0.09)`
- e) `1-pbinom(4,20,0.09)`

Spurning 9

Látum Y tákna fjölda þorpsbúa sem eru með sjúkdóminn þegar 20 þorpsbúar eru valdir að handahófi. Hvert er staðalfrávik Y ?

- a) 0.09.
- ☒ b) 1.28.
- c) 1.64.
- d) 1.8.
- e) 20.

Sigrún Helga biður nemendur sína sem og vini um að svara spurningakönnun. Í könnuninni er meðal annars spurningin: "Hversu forvitin(n) ertu á skalanum 0 - 100?". Niðurstöðurnar má sjá á myndinn hér að neðan:



Spurningar 10 og 11 fjalla um þessa könnun Sigrúnar.

Spurning 10

Hvert af eftirfarandi lýsir lögun dreifingar breytunnar best?

- a) Samhverf dreifing
- b) Samhverf fjölkryppudreifing
- c) Vinstri skekkt dreifing
- d) Hægri skekkt dreifing
- e) Samhverf tvíkryppudreifing

Spurning 11

Hver gætu verið mögulegt meðaltal og miðgildi mælinga breytunnar?

- a) Meðaltalið er 103 og miðgildið er 79.
- b) Meðaltalið er 79 og miðgildið er 83.
- c) Meðaltalið er 122 og miðgildið er 148.
- d) Meðaltalið er 23 og miðgildið er 42.
- e) Meðaltalið er 42 og miðgildið er 23.

Theódóra tölvunarfræðingur framkvæmdi tilraun til að kanna hvort samband væri milli kyns og hvort háskólanemendur nota Linux eða eitthvað annað stýrikerfi. Hún valdi 100 háskólanema af handahófi og spurði þá hvers kyns þeir væru og hvort þeir notuðu Linux. Niðurstöðurnar má sjá hér að neðan:

	Notar ekki Linux	Notar Linux
Karlar	43	9
Konur	44	4

Spurningar 12 og 13 fjalla um Theódóru og tilraunina hennar.

Spurning 12

Nú framkvæmdi Theódóra viðeigandi tilgátupróf (með $\alpha = 0.05$) til að kanna hvort samband sé á milli kyns og hvort háskólanemendur nota Linux. Hver af eftirfarandi fullyrðingum er sönn?:

- a) Við höfnum núlltilgátunni ef viðeigandi prófstærð er stærri en 1.96.
- b) Við höfnum núlltilgátunni ef viðeigandi prófstærð er stærri en 3.841.
- c) Við höfnum núlltilgátunni ef viðeigandi prófstærð er stærri en 5.024.
- d) Við höfnum núlltilgátunni ef viðeigandi prófstærð er stærri en 5.991.
- e) Við höfnum núlltilgátunni ef viðeigandi prófstærð er stærri en 7.378.

Spurning 13

Theódóra byrjaði á að búa til töflu sem sýnir væntanlega tíðni og var byrjuð á að búa til töflu sem sýnir framlag til prófstærðar. Hún var búin að reikna þrjú gildi en það fjórða vantar. Töfluna má sjá hér að neðan:

Framlag til prófstærðar	Notar ekki Linux	Notar Linux
Karlar	0.111	0.742
Konur	0.120	

Hvert er gildið á prófstærðinni sem Theódóra á að nota til að kanna hvort samband sé á milli kyns og hvort háskólanemendur nota Linux?

- a) 1.777
- b) 0.973
- c) 1.342
- d) 0.653
- e) 1.934

Magnús dýrafræðingur gerði rannsókn til að kanna mun á meðallengd kvenkyns og karlkyns húsamúsa. Gera má ráð fyrir að lengd kvenkyns og karlkyns húsamúsa fylgi normaldreifingu í báðum þýðunum. Magnús tók slembiúrtak tuttugu kvenkyns músa og tuttugu karlkyns músa og mældi lengdir þeirra allra. Mælingar hans sýndu að úrtaksmeðaltal kvenkyns músanna var 8.88cm með staðalfrávik 3.67cm á meðan úrtaksmeðaltal karlkyns músanna var 10.60cm með staðalfrávik 4.21cm. Köllum kvenkyns húsamýs hóp 1 og karlkyns húsamýs hóp 2.

Spurningar 14 - 16 fjalla um Magnús og tilraunina hans.

Spurning 14

Hvað af eftirfarandi er rétt miðað við gefnar upplýsingar?

- a) $\mu_1 = 8.88$
- b) $\sigma_1^2 = 17.7$
- c) $\mu_2 = 8.88$
- d) $\sigma_2^2 = 4.21$
- e) Ekkert af ofangreindu er rétt

Spurning 15

Magnús framkvæmdi nú viðeigandi tilgátupróf ($\alpha = 0.05$) til að kanna tilgátuna:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{á móti} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

P-gildi prófsins var 0.562. Byggt á þeim niðurstöðum, hvers konar tilgátupróf er viðeigandi að nota til að kanna hvort munur sé á meðallengd kvenkyns og karlkyns húsamúsa?

- a) Parað t-próf.
- b) Tveggja hópa t-próf, svipuð dreifni.
- c) Tveggja hópa t-próf, ólík dreifni.
- d) F-próf fyrir dreifni tveggja normaldreifðra þýða.
- e) Kí-kvaðratpróf fyrir dreifni normaldreifðs þýðis.

Spurning 16

Magnús framkvæmdi viðeigandi tilgátupróf til að kanna hvort munur sé á meðallengd kvenkyns og karlkyns húsamúsa og fékk út p-gildið 0.176. Gerum nú ráð fyrir að í raun sé munur á meðallengd kvenkyns og karlkyns húsamúsa. Hver eftirfarandi fullyrðinga er sönn ($\alpha = 0.05$)?

- a) Villa af gerð I átti sér stað.
- b) Villa af gerð II átti sér stað.
- c) Bæði villa af gerð I og villa af gerð II áttu sér stað.
- d) Hvorki villa af gerð I né villa af gerð II áttu sér stað.
- e) Engin af fullyrðingum hér að ofan er rétt.

Gunnar tölfræðingur hefur mikinn áhuga á húsnæðisverði í ákveðnu hverfi í Reykjavík. Hann safnaði gögnum um verð (nuvirdi) og stærð (ibm2) á 50 íbúðum í hverfinu með það að markmiði að smíða líkan til að spá fyrir um verð íbúða með stærð. Hann notað R til að smíða líkan og má sjá hluta af R-úttaki hér að neðan.

Call:

```
lm(formula = nuvirdi ~ ibm2, data = ahj)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8657690	2476749	3.496	0.00103 **
ibm2	208498	31030	6.719	0.0000000198 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5235000 on 48 degrees of freedom

Multiple R-squared: 0.4847, Adjusted R-squared: 0.4739

F-statistic: 45.15 on 1 and 48 DF, p-value: 0.00000001983

Spurningar 17 og 18 fjalla um Gunnar og líkanið hans.

Spurning 17

Hver er aðhvarfslína minnstu kvaðrata fyrir líkanið hans Gunnars?

- a) $\hat{y} = 208498 + 8657690 \cdot x$
- b) $\hat{y} = 3.496 + 6.719 \cdot x$
- c) $\hat{y} = 8657690 + 2476749 \cdot x$
- d) $\hat{y} = 208498 + 31030 \cdot x$
- e) $\hat{y} = 8657690 + 208498 \cdot x$

Spurning 18

Hver er fylgnistuðullinn, r , milli verðs og stærðar íbúða?

- a) 0.4874
- b) 0.6962
- c) 0.2376
- d) 0.8734
- e) Ekki er hægt að finna gildið á fylgnistuðlinum með þeim upplýsingum sem gefnar eru.

Spurning 19

Elínborg efnafræðingur framkvæmdi rannsókn til að bera saman meðaltöl tveggja þýða. Hún reiknaði 95% öryggisbil fyrir mismun meðaltala þýðanna sem: $-2.31 < \mu_1 - \mu_2 < 4.21$. Hún hefur áhuga á að kanna hvort munur sé á meðaltölum þýðanna og ætlar hún að nota $\alpha = 0.05$. Hún framkvæmir viðeigandi tilgátupróf. Hver af eftirfarandi fullyrðingum er sönn?

- a) p-gildi prófsins er hærra en 0.05.
- b) Líkurnar á að núlltilgátan sé sönn eru 0.05.
- c) Líkurnar á að gagntilgátan sé sönn eru 0.05.
- d) Elínborg getur hafnað núlltilgátunni.
- e) Engin af fullyrðingunum hér að ofan er rétt.

Spurning 20

Íris spurði 280 einstaklinga hvort þeir eiga maka og hvort þeir eigi barn. Niðurstöðurnar má sjá hér að neðan.

```
> table(dat$kaerast.factor, dat$barn.factor)
```

	á barn	á ekki barn
á maka	133	89
á ekki maka	13	45

Hver eftirfarinna fullyrðinga er rétt miðað við niðurstöðurnar hér að ofan?

- a) 66% þeirra sem eiga maka, eiga líka barn.
- b) 66% þeirra sem eiga barn, eiga líka maka.
- c) 91% þeirra sem eiga barn, eiga líka maka.
- d) 91% þeirra sem eiga maka, eiga líka barn.
- e) 5% þeirra sem eiga ekki maka, eiga barn.

Krossaspurningasvarblað: Merkið með X við það svar sem þið teljið vera það rétta í krossaspurningunum sem eru á bls. 9-16.

Spurning 1	a	b	c	d	e
Spurning 2	a	b	c	d	e
Spurning 3	a	b	c	d	e
Spurning 4	a	b	c	d	e
Spurning 5	a	b	c	d	e
Spurning 6	a	b	c	d	e
Spurning 7	a	b	c	d	e
Spurning 8	a	b	c	d	e
Spurning 9	a	b	c	d	e
Spurning 10	a	b	c	d	e
Spurning 11	a	b	c	d	e
Spurning 12	a	b	c	d	e
Spurning 13	a	b	c	d	e
Spurning 14	a	b	c	d	e
Spurning 15	a	b	c	d	e
Spurning 16	a	b	c	d	e
Spurning 17	a	b	c	d	e
Spurning 18	a	b	c	d	e
Spurning 19	a	b	c	d	e
Spurning 20	a	b	c	d	e

Takk fyrir veturinn og hafið það sem allra best í sumar!