# Experimentálna ekonómia

## Prednáška 6: Deskriptívna štatistika

Matej Lorko
matej.lorko@euba.sk
Materiály: www.lorko.sk/lectures

Referencie:


- Weimann, J., & Brosig-Koch, J. (2019). Methods in experimental economics. Springer International Publishing. Chicago
- Jacquemet, N., & l'Haridon, O. (2018). Experimental economics. Cambridge University Press.
- Rajčáni, J., Kačmár, P., Bavolar, J., Cavojova, V., Adamkovič, M., Vargová, L., Kočišová, L., Martončik, M. (2024). Štatistika pre reprodukovateľný výskum v spoločenských vedách.

# Výber vhodnej štatistickej metódy analýzy

- Aký je účel štatistickej analýzy: poskytnúť popisnú reprezentáciu dát a efektov treatmentu? Urobiť štatistický záver týkajúci sa populácie, z ktorej sa vzorka čerpá (inferencia)? Robiť predpoveď na základe odhadovaného modelu? Aké sú hlavné štatistické charakteristiky experimentálneho dizajnu alebo výsledných dát (odpovede z predchádzajúcich otázok)? Aké analytické metódy možno použiť vzhľadom na hlavné štatistické charakteristiky?

- Spracovanie dát: Chýbajú v dátach niektoré hodnoty? Existujú odľahlé pozorovania (outliers)? Existujú subjekty, ktoré zjavne urobili náhodné rozhodnutia (nepochopili experiment)?

- Vytváranie nových premenných z (kombinácie) už zhromaždených premenných (napr. skupinové priemery); Vytvorenie zoznamu premenných s popisom.

- Analýza dát: Popis dát pomocou kľúčových ukazovateľov; Grafické znázornenie dát; prispôsobenie štatistického modelu dátam a odhadom parametrov modelu; Diagnostika modelu; Vyvodzovanie záverov; Predpovede pomocou odhadovaného modelu.

- Závery: Dajú sa účinky treatmentu štatisticky overiť? Dokáže model dobre vysvetliť pozorované dáta? Sú potrebné ďalšie experimentálne treatmenty?

# Popis dát

- Najlepším spôsobom, ako sa naučiť písať o dátach, je prečítať si dátové sekcie v iných článkoch a venovať sa druhu informácií, ktoré obsahujú. Vaša sekcia s dátami by mala obsahovať minimálne toto:

- Identifikujte zdroj dát. Zahrňte vetu, ktorá explictine hovorí, odkiaľ vaše dáta pochádzajú.

- Popíšte dáta. Mali by ste napríklad uviesť počet pozorovaní, počet vzoriek, časové obdobie, počas ktorého sa dáta zhromažďovali, spôsob zberu dát atď.

- Uveďte silné a slabé stránky dát. Aké sú vaše dáta v porovnaní s inými dátami použitými v literatúre? Poskytuje váš výskum viac pozorovaní alebo novších pozorovaní než iné výskumy? Boli údaje zhromaždené spoľahlivejším spôsobom? Prečo sú tieto dáta pre vašu štúdiu obzvlášť vhodné (alebo nie)? Všimnite si všetky vlastnosti dát, ktoré môžu ovplyvniť vaše výsledky. Boli určité skupiny populácie nadmerne alebo nedostatočne zastúpené? Existuje skreslenie spôsobené prirodzeným úbytkom (attrition) alebo samo-výberom (selection bias)? Zmenil sa počas zhromažďovania dát spôsob merania?

- Vysvetlite všetky vykonané výpočty alebo úpravy. Dáta vám niekedy niečo neposkytú priamo; možno ste museli pridať / odčítať / vynásobiť / rozdeliť dva zadané údaje, aby ste získali tretí. Popíšte, ako ste vybrali svoju vzorku. Museli ste napríklad vylúčiť určité druhy pozorovaní?

# Operacionalizácia premenných

- Výsledkom merania je súbor obsahujúci určité premenné, čiže vlastnosti, ktoré môžu nadobúdať rôzne hodnoty. Pri meraní sú dôležité pravidlá, na základe ktorých číselné hodnoty priraďujeme – teda ich operacionalizácia. Nemôžeme sa spoliehať na to, že vždy existuje iba jeden správny spôsob operacionalizácie premenných. Práve naopak, je na nás, ako premenné uchopíme – závisí to aj od našich výskumných východísk a témy, ako presne chceme určité premenné merať.

- Operacionalizácia musí byť konzistentná – teda nemôžeme u jedného participanta merať vek v rokoch a u iného v mesiacoch a pod.

- Musíme byť úplne transparentní čo sa týka operacionalizácie premenných. Častým problémom totiž je, že na základe opisu výskumu v študentskej práci alebo v článku, čitatelia nie sú schopní zistiť, ako sú premenné operacionalizované. Väčšinou je to nedopatrením zo strany výskumníkov. V bežnom živote sa skrátka príliš často spoliehame na to, že druhí automaticky pochopia to, čo sme mysleli. Že je to samozrejmé, jasné z kontextu a pod. Na naše prekvapenie je toto očakávanie aj v bežnom živote často úplne mylné – nedorozumenia vznikajú na dennom poriadku. Vo vede však musíme nedorozumeniam predchádzať. Nič teda nie je samozrejmé. Všetky informácie musíme jasne a transparentne uviesť.

- Pri operacionalizácii premennej je nevyhnutné reflektovať, na akej škále premennú meriame – keďže rovnakú premennú môžeme merať na rôznych škálach.

# Typy premenných

- Nominálna premenná (tiež kategorická, resp. kvalitatívna) nevyjadruje kvantitu. Ide iba o číselné označenie určitého javu. Príkladom môže byť farba očí, ktorá môže byť napr. hnedá (môžeme ju kódovať ako („1“); zelená („2“), modrá („3“); alebo iná („4“). Ide však o samostatné kvality, ktoré nijak nesúvisia s ich číselným označením. To, že je modrá označená číslom „3“ a zelená „2“, neznamená, že modrá má niečoho viac alebo menej.

- Poradová premenná (tiež ordinálna) na rozdiel od nominálnej už vyjadruje určitú základnú mieru kvantity, teda určité poradie. Ak by sme napríklad zoradili všetky študentky a študentov v triede podľa výšky vzostupne, vieme povedať, že osoba s vyšším číslom je vyššia. Typickým príkladom poradovej škály môže byť jedna sebavýpoveďová škála, napr.: Nakoľko ste spokojný/spokojná s kvalitou tohto predmetu? 1 = Vôbec nie som spokojný/spokojný, 2 = Skôr nie som spokojný/spokojný, 3 = Nie som ani spokojný/spokojná, ani nespokojný/nespokojný, 4 = Skôr som spokojný/spokojná, 5 = Som veľmi spokojný/spokojná

- Problémom poradových škál je však skutočnosť, že pri poradí nevieme povedať, aký veľký je rozdiel, teda o koľko sa jedna úroveň líši od inej? Pri poradovej premennej vieme premenné usporiadať, ale nie sčítavať, odčítavať násobiť ani deliť.

- Kvantitatívna premenná (tiež kardinálna) je na rozdiel od nominálnej a poradovej skutočným číslom. Rozdiely medzi jednotlivými bodmi na škále sú zmysluplné, môžeme povedať, že premenná má jednotku merania. Kvantitatívne premenné rozlišujeme na dva typy – intervalové a pomerové.

  - Intervalové premenné majú štandardnú jednotku, nemajú však pravý nulový bod, ktorý by označoval úplnú absenciu danej kvality. Najbežnejším príkladom je teplota v stupňoch Celzia (°C). Ak je dnes 20 °C a včera bolo 23 °C, rozdiel medzi týmito hodnotami je 3 °C, ide teda o rovnaký rozdiel ako medzi 17° C a 20 °C. Na druhej strane 0°C neznamená absenciu akejkoľvek teploty, je to iba dohodnutá hodnota zodpovedajúca teplote topenia vody. Ak je teda dnes 20 °C a minulý týždeň bolo 10 °C, bolo by zvláštne tvrdiť, že dnes je dvakrát teplejšie. Numericky to dáva zmysel, obsahovo ale úplne nie. Sčítavanie a odčítavanie intervalových premenných problémy nerobí, no násobenie a delenie už môže.

  - Druhým typom kvantitatívnej premennej je pomerová premenná (ratio), ktorá predstavuje skutočné číslo s pravým nulovým bodom. Príkladom môže byť počet (napr. počet bodov v teste), dĺžka meraná v metroch, reakčný čas v milisekundách, alebo srdcový tep meraný v počte úderov za minútu. 0 pri nich nie je dohodnutá vec ale absencia daného javu (bodov, dĺžky, trvania, úderov srdca). S pomerovou premennou vieme robiť všetky matematické operácie.

# Miery centrálnej tendencie

- Ak chceme celý súbor alebo populáciu charakterizovať jedným údajom, užitočné je opísať tzv. centrálnu tendenciu (alebo stred rozdelenia). Okrem aritmetického priemeru centrálnu tendenciu vyjadrujú aj modus a medián.

- Aritmetický priemer (mean, $\bar{X}$ , M) premennej X (X reprezentuje vo vzorci názov premennej) vypočítame jednoduchým súčtom všetkých prvkov v súbore, ktorý následne vydelíme počtom prvkov.  Priemer má svoje predpoklady a špecifiká. Môžeme ho použiť iba pri kvantitatívnych dátach, keďže priemer počítame prostredníctvom súčtu a podielu. Jedným z limitov priemeru je jeho citlivosť na extrémne skóre, teda také údaje, ktoré sa veľmi výrazne odlišujú od zvyšku súboru. Výskyt extrémnych dát môže viesť k skresleniu priemeru tak, že prestáva dobre opisovať stred dát. Typickým príkladom premennej, kde je priemer ovplyvnený extrémnymi hodnotami, je finančný príjem (mzda) v populácii. Ak sa pokúšame porovnať našu mzdu s ostatnými ľuďmi v krajine, údaj o priemernej mzde nám nebude veľmi užitočný, keďže pomerne malý počet vysokopríjmových jednotlivcov skresľuje priemer tak, že jeho hodnota je vyššia, ako je stredná hodnota v populácii. Ak by sme teda uvažovali o priemernej mzde ako o strednej mzde človeka v tejto krajine, mýlili by sme sa. Tu by nám bol užitočnejší medián.

- Medián ($\tilde{X}$, Md) je stredná hodnota dát v súbore. Ak všetky prvky v súbore usporiadame podľa veľkosti, medián bude presne v strede a bude rozdeľovať súbor na dve rovnako početné polovice. Inak povedané, polovica prípadov bude mať hodnotu vyššiu ako medián a polovica prípadov bude mať hodnotu nižšiu, ako je medián. Ak teda zistím, že moja mzda je vyššia ako medián, viem, že väčšina ľudí v krajine zarába menej.  Ak by sme teda mali súbor s hodnotami: 12, 13, 15, 18, 21, mediánom je hodnota 15. V prípade, že máme párny počet prvkov, v súbore nebudeme mať jednu strednú hodnotu, ale dve. V tom prípade je medián ich priemerom. Napríklad v súbore s hodnotami: 11, 13, 14, 15, 16, 21 je medián 14,5. Medián je vhodnejší na prezentáciu, keď pracujeme s poradovými dátami. Zároveň, na rozdiel od priemeru, medián nie je citlivý na extrémne dáta, preto môže byť vhodnejšou mierou stredu v prípade, že pracujeme s dátami, ktoré extrémy obsahujú. Na druhej strane medián môže podceňovať význam dôležitých hodnôt na krajoch rozdelenia.

- Modus (Mo) je relatívne jednoduchou štatistikou. Ide o najčastejšie sa vyskytujúcu hodnotu v súbore. Je možné uplatniť ho na akékoľvek dáta vrátane kvalitatívnych. V niektorých súboroch môžeme získať viac modálnych hodnôt, označujeme ich ako multimodálne (bimodálne v prípade dvoch modov). Treba si však uvedomiť, že informačná hodnota modu je obmedzená, pretože nijako nereflektuje rozdelenie ostatných hodnôt, a tak sa môže stať, že modus nebude dobrým reprezentantom celého súboru.

# Miery variability

- Štatistiky, s ktorými sme doteraz pracovali, slúžili na charakteristiku rozdelenia na základe jeho stredu. Dôležitú informáciu nám poskytujú aj miery variability. Tie odpovedajú na otázku, nakoľko sa dáta od seba líšia, resp. ako ďaleko od stredu sa naše dáta vyskytujú.

- Rozsah (rozpätie, range) premennej vyjadruje rozdiel najvyššieho a najnižšieho dosiahnutého skóre. Samostatne môžeme dáta opísať tak, že uvedieme minimum a maximum. V tomto prípade ide o veľmi jednoduché opísanie variability, avšak môže byť veľmi nepresné, keďže rozsah zahŕňa aj veľmi vzdialené hodnoty. Navyše rozsah nezachytáva relatívne početnosti jednotlivých hodnôt, nevieme teda povedať, aké časté sú ktoré hodnoty.

- Medzikvartilový rozsah (interquartile range – IQR) predstavuje rozsah stredných 50 % všetkých dát, čo je rozsah medzi prvým a tretím kvartilom (alebo 25. a 75. percentilom). Ak by sme dáta rozdelili podľa veľkosti na štyri rovnako početné časti, získame hodnoty ohraničujúce kvartily. Ak by sme ich usporiadali podľa veľkosti a rozdelili na 100 častí, získame ohraničenia percentilov. Hranica 1. a 2. kvartilu, ktorá predstavuje 25. percentil, ohraničuje 25 % najnižších hodnôt v súbore. 50. percentil je medián, keďže ide o hodnotu, ktorá rozdeľuje dáta na dve rovnako početné polovice. 75. percentil je hranicou medzi tretím a štvrtým kvartilom, ohraničuje nám tak 25 % najvyšších hodnôt.

- Štandardná odchýlka (SD, s) je asi najbežnejším ukazovateľom variability, ktorý sa zvykne prezentovať spolu s priemerom. Ak pracujeme s kvantitatívnymi dátami (intervalové, pomerové), môžeme vyjadriť variabilitu na základe odchýlok od priemeru. Tie môžeme vypočítať ako rozdiel každej hodnoty v súbore a priemeru. Keďže niektoré odchýlky sú pozitívne a niektoré negatívne, sčítaním by sa vynulovali. Budeme teda pracovať so štvorcami (druhými mocninami) odchýlok, ktoré už sčítať môžeme.

- Súčet štvorcov odchýlok nám síce vyjadruje celkovú variabilitu, je však závislý od počtu prvkov v súbore (keďže ide o súčet). Aby sme ho štandardizovali, je potrebné vydeliť ho počtom stupňov voľnosti (df), čo je počet pozorovaní mínus 1 (N – 1). Tým získame rozptyl (variancia, s2, σ2).

- Keďže interpretačne nemá veľmi zmysel rozprávať sa o štvorcoch odchýlok, je vhodné varianciu opäť odmocniť. Tým získame štandardnú odchýlku, ktorá je opäť vyjadrená v pôvodných jednotkách. Štandardná odchýlka nám teda hovorí o tom, o koľko sa nám naše dáta „v priemere odchyľujú od priemeru“.

- Symboly SD a s vyjadrujú štandardnú odchýlku odhadovanú na základe vzorky. Symbol σ predstavuje hypotetickú štandardnú odchýlku v populácii. Vo väčšine prípadov ju presne nepoznáme, keďže nemáme dáta z celej populácie, iba sa jej približujeme. Rovnako je to aj so symbolmi označujúcimi rozptyl (s2, σ2).

# Náhodné premenné a ich distribúcia

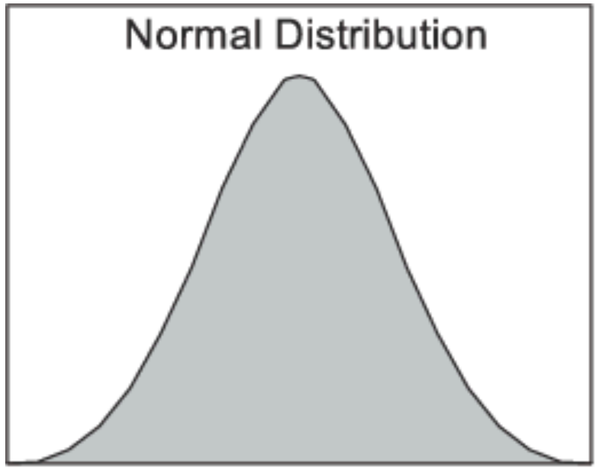
- Pri štatistickom modelovaní vzťahu medzi premennými sa závislá premenná interpretuje ako náhodná premenná. Hodnoty náhodnej premennej ktoré sú najpravdepodobnejšie a tie, ktoré sú menej pravdepodobné, určuje jej rozdelenie. Funkcia hustoty diskrétnej náhodnej premennej udáva pravdepodobnosť, s akou nastane určitá hodnota. Napríklad výsledky hodu kockou sú rovnomerne rozdelené, každý s príslušnou pravdepodobnosťou 1/6.

- V prípade spojitej náhodnej premennej, ako je dĺžka času rozhodnutia, nie je možné určiť pravdepodobnosť individuálnej hodnoty. Ak existuje nekonečný počet hodnôt, pravdepodobnosť jednej hodnoty musí byť nekonečne blízka nule. Z tohto dôvodu je pri spojitých premenných možné indikovať špecifické pravdepodobnosti len pre rozsahy hodnôt, pričom celková plocha pod funkciou hustoty je vždy 1. Kumulatívna (spojitá) distribučná funkcia je, matematicky povedané, integrálom spojitej funkcie hustoty. Hodnota funkcie v bode x teda udáva pravdepodobnosť, s ktorou náhodná premenná nadobudne hodnotu menšiu alebo rovnú x.

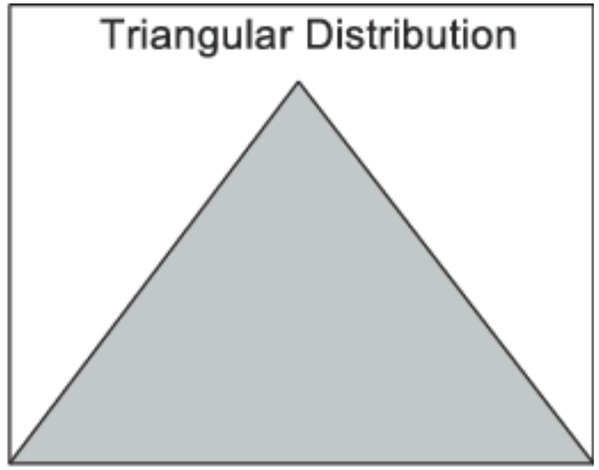
- Väčšina štatistických rozdelení má určité parametre, ktoré určujú tvar funkcie hustoty. Tri najdôležitejšie parametre sú očakávaná hodnota, rozptyl a stupne voľnosti. Očakávaná hodnota je priemerom všetkých hodnôt, ak náhodnú vzorku (teoreticky) vytiahneme nekonečne veľa krát. Napríklad, keďže existuje rovnaká pravdepodobnosť hodu každého čísla na (normálnej) kocke, očakávaná hodnota je $1/6 \cdot (1 + 2 + 3 + 4 + 5 + 6) = 3{,}5$. Rozptyl je stredná kvadratická odchýlka všetkých realizácií očakávanej hodnoty a teda predstavuje informáciu o “rozhodení” náhodnej premennej. Čím je rozptyl väčší, tým je funkcia hustoty širšia a viac plochá.

- Matkou všetkých distribúcií je normálne rozdelenie. Jeho parametrami sú očakávaná hodnota $\mu$ a rozptyl $\sigma 2$. Hustota pravdepodobnosti je zvonovitá a symetrická okolo $\mu$, kde má najvyššiu funkčnú hodnotu hustoty. Iné dôležité rozdelenia nie sú parametrizované priamo pomocou očakávanej hodnoty a rozptylu, ale aj nepriamo pomocou takzvaných stupňov voľnosti, ktoré ovplyvňujú očakávanú hodnotu a/alebo rozptyl. Napríklad (Studentovo) t-rozdelenie má takéto stupne voľnosti, pričom tvar funkcie hustoty sa stále viac a viac približuje funkcii hustoty štandardného normálneho rozdelenia so zvyšujúcimi sa stupňami voľnosti.

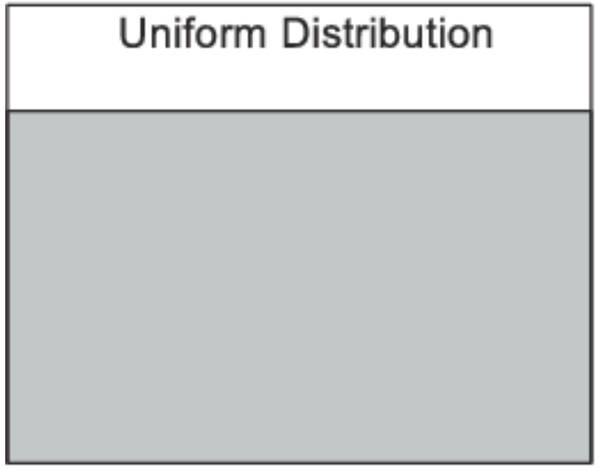
# Náhodné premenné a ich distribúcia



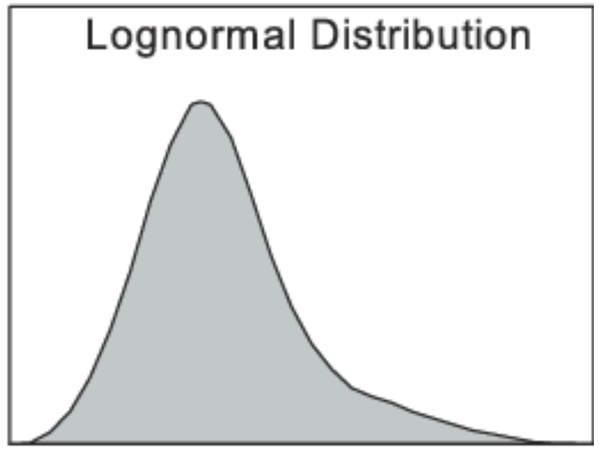

Normal distribution




Triangular distribution




Uniform distribution




Log-normal distribution

# Deskriptívne štatistiky

- Dátove sekcie v článkoch často obsahujú tabuľku deskriptívnych štatistík - teda štatistík ktoré určujú relevantnosť vzorky. Zvyčajne zahŕňajú priemer (napr. priemerný príjem, priemerný vek, priemerné roky školskej dochádzky atď.) a štandardnú odchýlku. V prípade kategorických údajov (napríklad rasy) však neuvádzate priemer; namiesto toho použijete percento pozorovaní v každej skupine.

- Očakávaná hodnota - Priemerná hodnota atribútu vzorky, založená na opakovanom výbere vzoriek z populácie.

- Štandardné chyby - smerodajná odchýlka alebo miera variability / rozptylu vo vzorke. Čím väčšia je vzorka, tým menšia je štandardná chyba.

- Distribúcie vo vzorke - Teoretické (nepozorované) rozdelenie atribútu, ktoré umožňuje výpočet intervalov spoľahlivosti a testy hypotéz.

- POZNÁMKA: Priemer a štandardná odchýlka fungujú dobre pre normálne rozdelenie (v tvare zvonovej krivky). Ak natrafíme na inú distribúciu, môže byť užitočnejšie použiť na opis centrálnej tendencie (očakávanej hodnoty) medián alebo modus.

# Grafické znázornenie dát

- Dobre zostavený graf dokáže odpovedať na niekoľko otázok naraz:
- Centrálna tendencia: Kde leží stred distribúcie?
- Rozptyl alebo variácia: Ako veľmi sú pozorovania roztiahnuté alebo koncentrované?
- Tvar distribúcie: Má distribúcia iba jediný vrchol (jedna koncentrácia pozorovaní v relatívne úzkom rozmedzí hodnôt), alebo má vrcholov viac?
- Chvosty: Približne aké percento pozorovaní leží na koncoch (chvostoch) distribúcie?
- Symetria alebo asymetria (nazývaná tiež nesúmernosť): Majú pozorovania tendenciu hromadiť sa na jednej strane distribúcie, zatiaľ čo na druhej ich je relatívne málo? Alebo má každá strana distribúcie zhruba rovnaký počet pozorovaní?
- Odľahlé hodnoty (outliers): Existujú hodnoty, ktoré sa v porovnaní s väčšinou javia ako veľmi veľké alebo veľmi malé?
- Porovnanie: Ako sa dve distribúcie líšia z hľadiska tvaru, natiahnutia a centrálnej tendencie?
- Vzťahy: Je možné, že hodnoty jednej premennej súvisia s hodnotami inej?

# Výber správneho grafu

**TABLE 11-10** Typical Presentation and Exploratory Graphs

| Type of Graph | What Is Displayed | Most Appropriate Level of Measurement | Number of Cases | Comments |
|---|---|---|---|---|
| Bar chart | Relative frequencies (percentages, proportions) | Categorical (nominal, ordinal) | 3–10 categories | Common presentation graphic |
| Dot chart | Frequencies, distribution shape, outliers | Quantitative (interval, ratio) | *Less than* 50 cases | Displays actual data values |
| Histogram | Distribution shape | Quantitative | $N > 50$ cases | Essential exploratory graph for interval or ratio variables with a large number of cases |
| Boxplot | Distribution shape, summary statistics, outliers | Quantitative | $N > 50$ cases | Can display several distributions; actual data points, an essential exploratory tool |
| Time series plot | Trends | Quantitative (percentages, rates) | $10 < N < 100$ | Common in presentation and exploratory graphics |

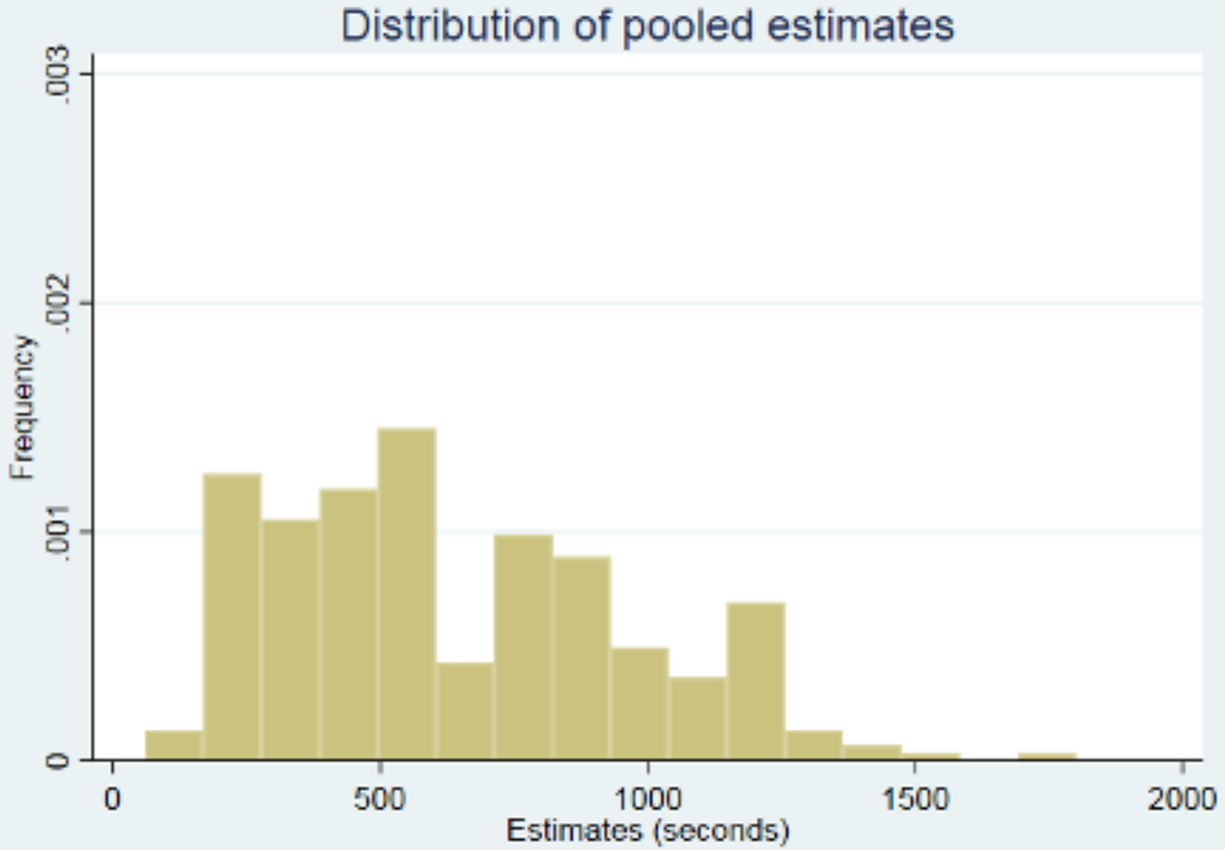
# Histogram

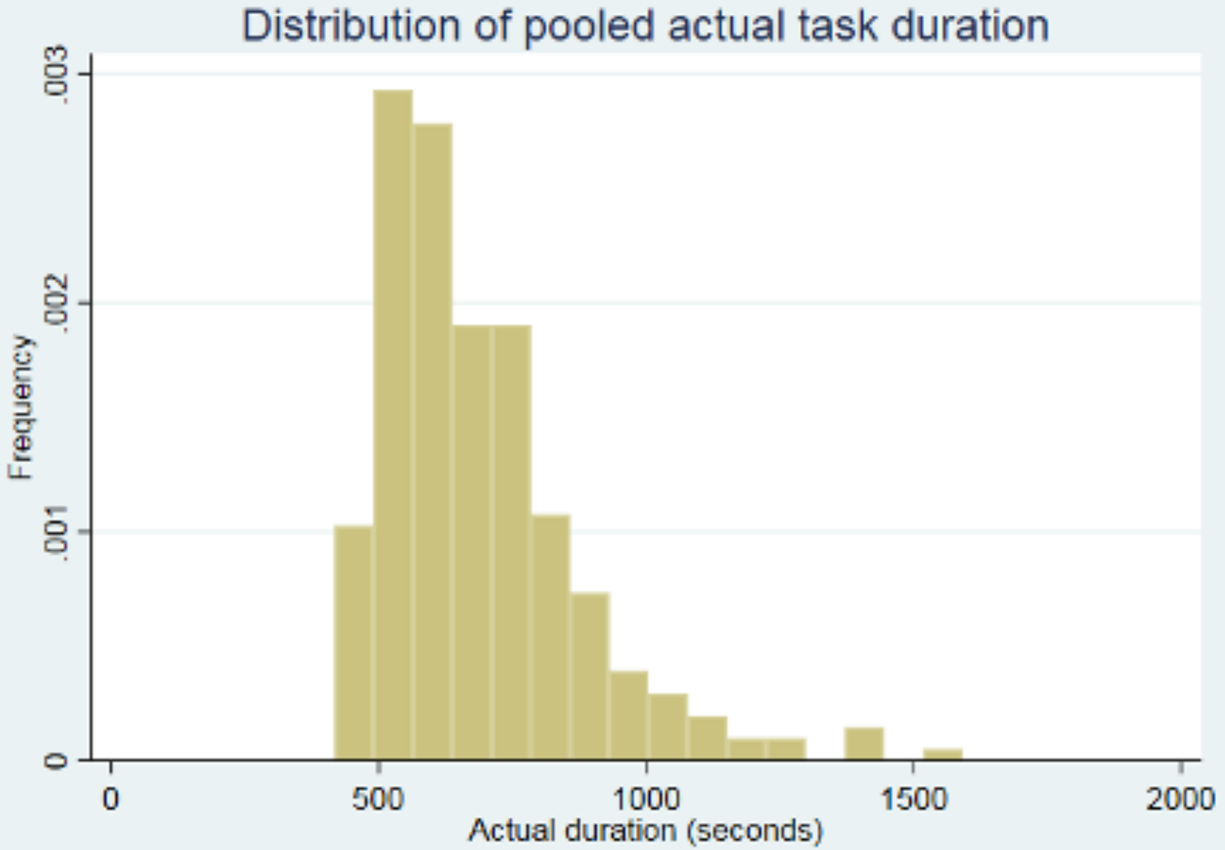
- V prípade, že pracujeme s kvantitatívnymi premennými, sa namiesto zobrazovania početností jednotlivých nameraných hodnôt snažíme skôr zobraziť tvar rozdelenia a dát prostredníctvom početností v určitých intervaloch hodnôt. K tomu nám slúžia histogramy. V histograme sa na osi x rozdelia všetky možné hodnoty do intervalov a výška stĺpca následne ukazuje početnosť (prípadne relatívnu početnosť) nameraných hodnôt v danom intervale.

- Histogram nám tak dáva dobrú predstavu o celkovom rozdelení dát. V rámci rozdelenia dát môžeme hodnotiť jeho symetrickosť zhustenie, zhluky, medzery (chýbajúce dáta), výskyt odľahlých hodnôt a tvar rozdelenia.

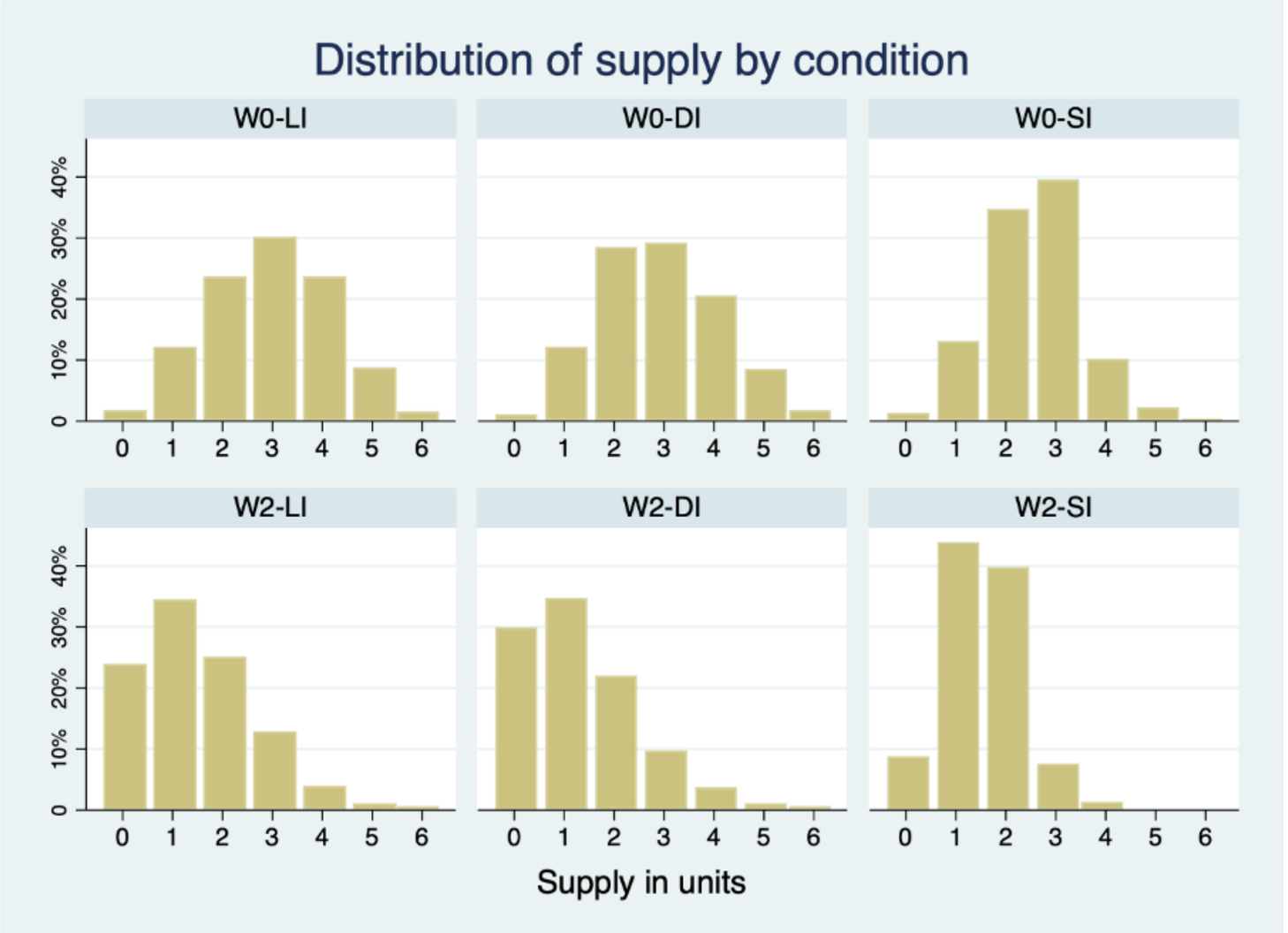
- Tvar rozdelenia opisujeme aj prostredníctvom číselných mier šikmosti a špicatosti. Taktiež si na histograme môžeme všimnúť počet vrcholov (najviac sa vyskytujúci interval hodnôt). Aj keď najčastejšie očakávame jeden vrchol (unimodálny tvar), môže sa stať, že vrcholy budú dva (bimodálny tvar) alebo viac ako dva (multimodálny tvar). Tvar histogramu s dvomi vrcholmi (samozrejme, jasne odlíšiteľné, ako keby sme mali dva zvonovité tvary vedľa seba) vzniká napríklad vtedy, ak údaje pochádzajú z dvoch rôznych rozdelení (napr. nejde o jednu, ale o dve rôzne populácie), a preto je potrebné tieto odlišné rozdelenia identifikovať a vykresliť ich samostatne.

# Histogram

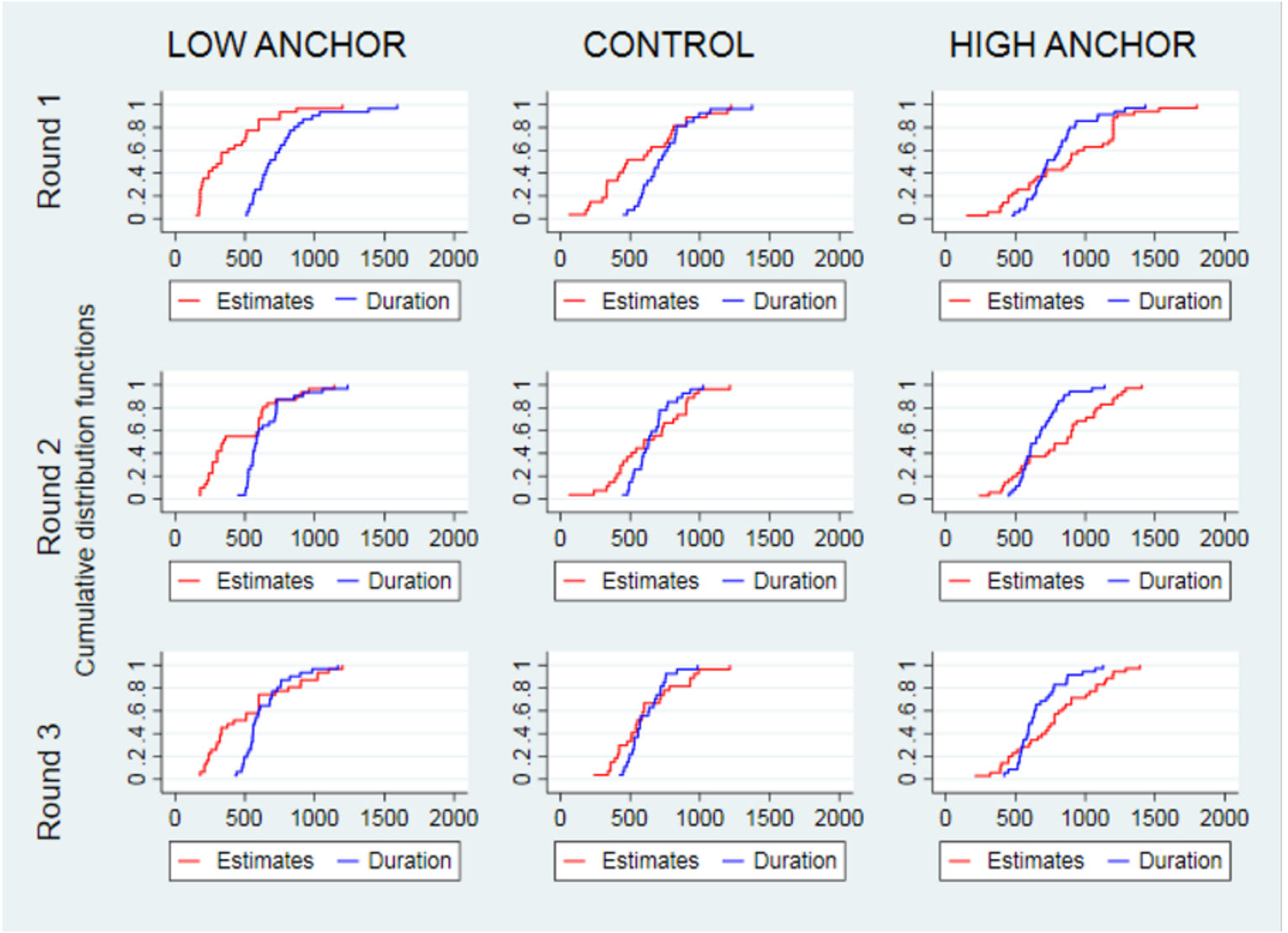

Distribution of pooled estimates
Frequency
0
.001
.002
.003
0
500
1000
1500
2000
Estimates (seconds)


Distribution of pooled actual task duration
Frequency
0
.001
.002
.003
0
500
1000
1500
2000
Actual duration (seconds)

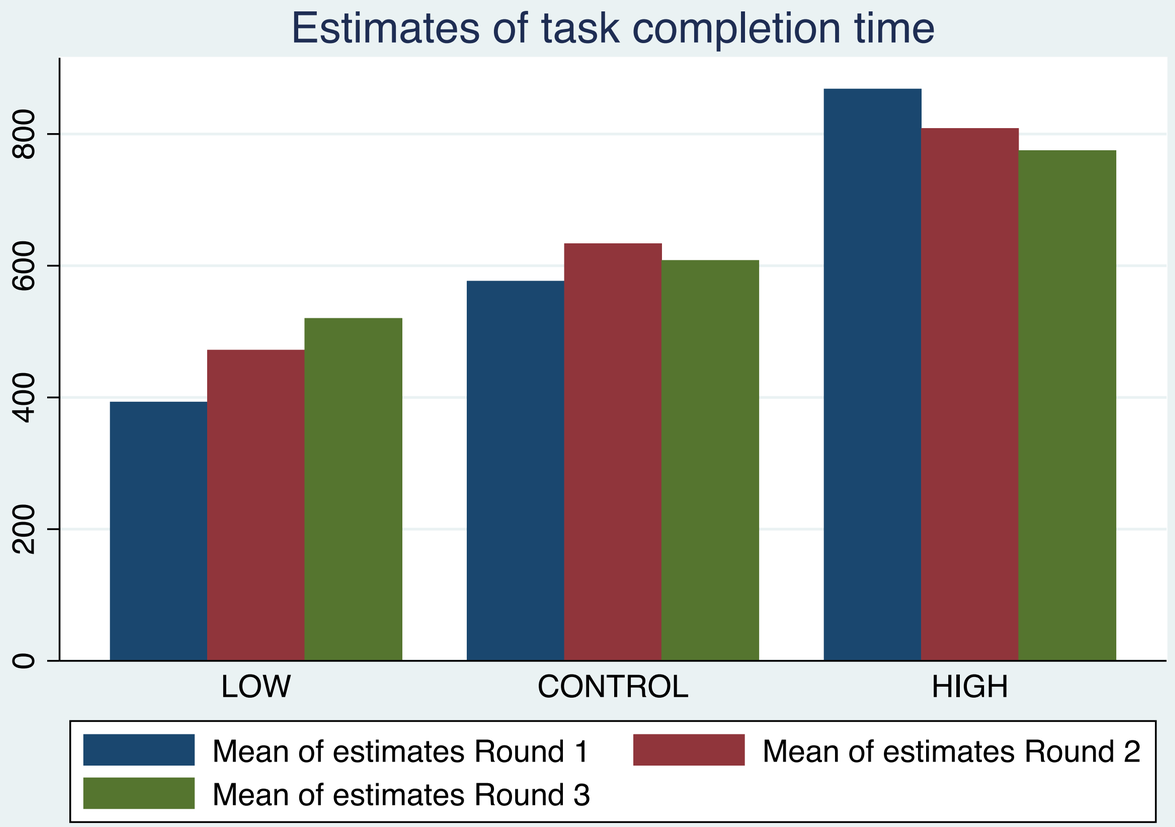
# Histogram

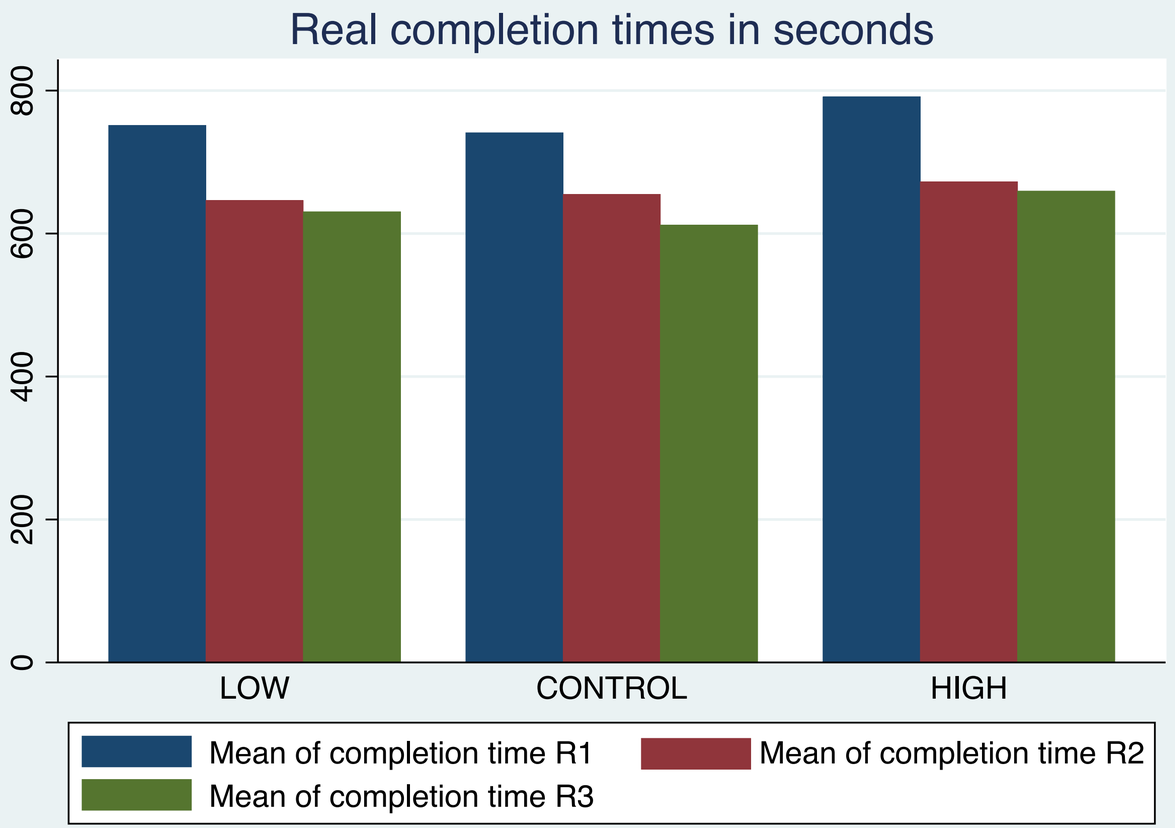

Distribution of supply by condition
W0-LI
W0-DI
W0-SI
W2-LI
W2-DI
W2-SI
0
10%
20%
30%
40%
0 1 2 3 4 5 6
Supply in units

# Kumulatívna distribúcia

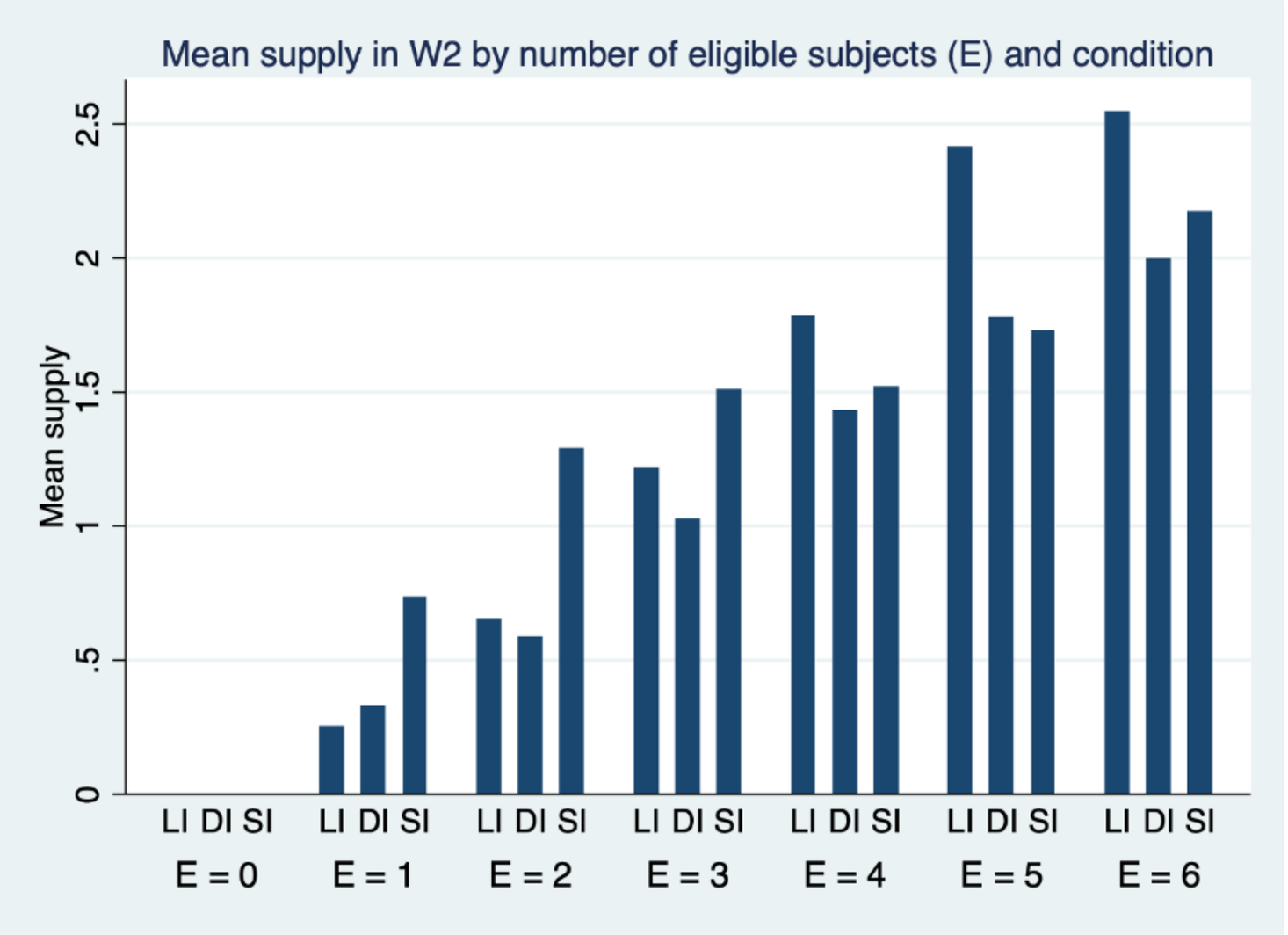


# Stĺpcový graf

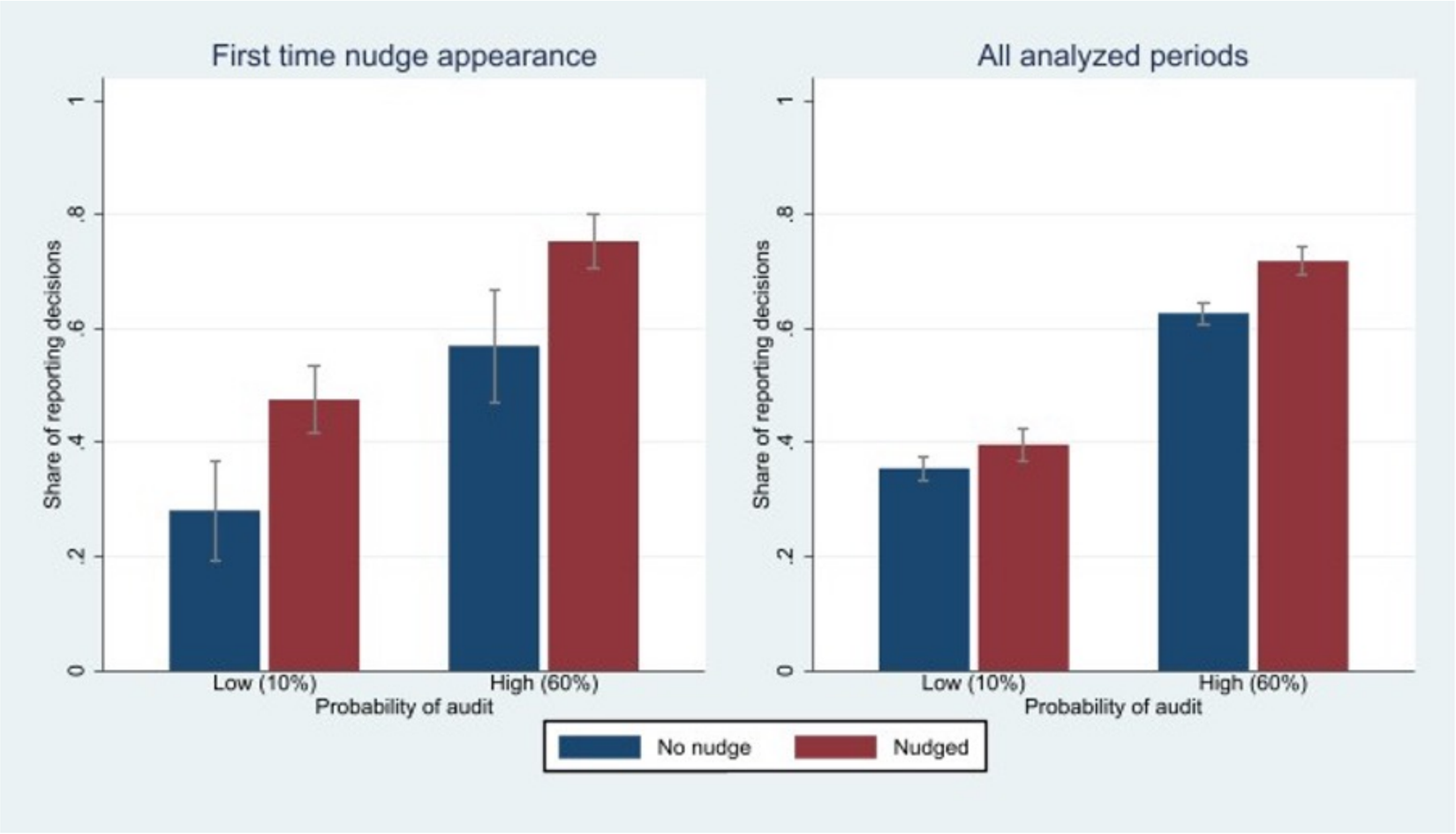


# Stĺpcový graf



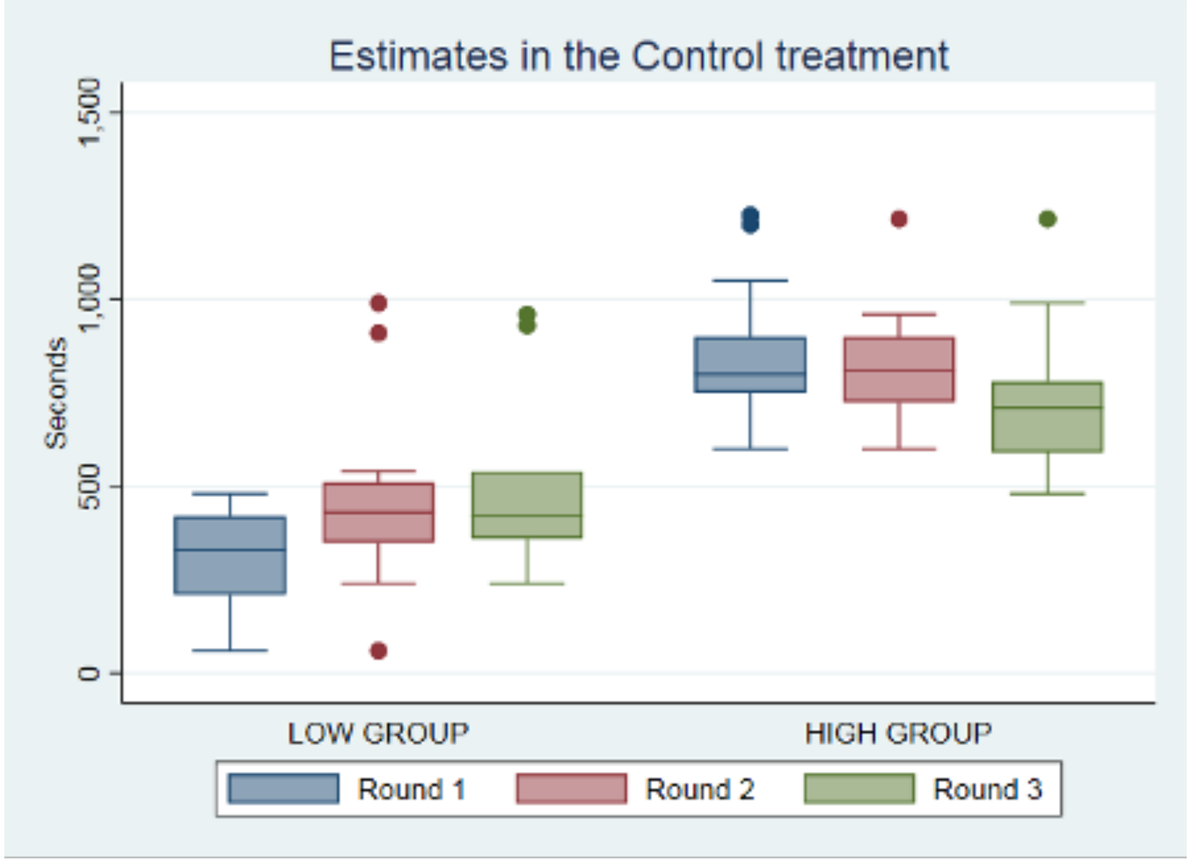
# Stĺpcový graf

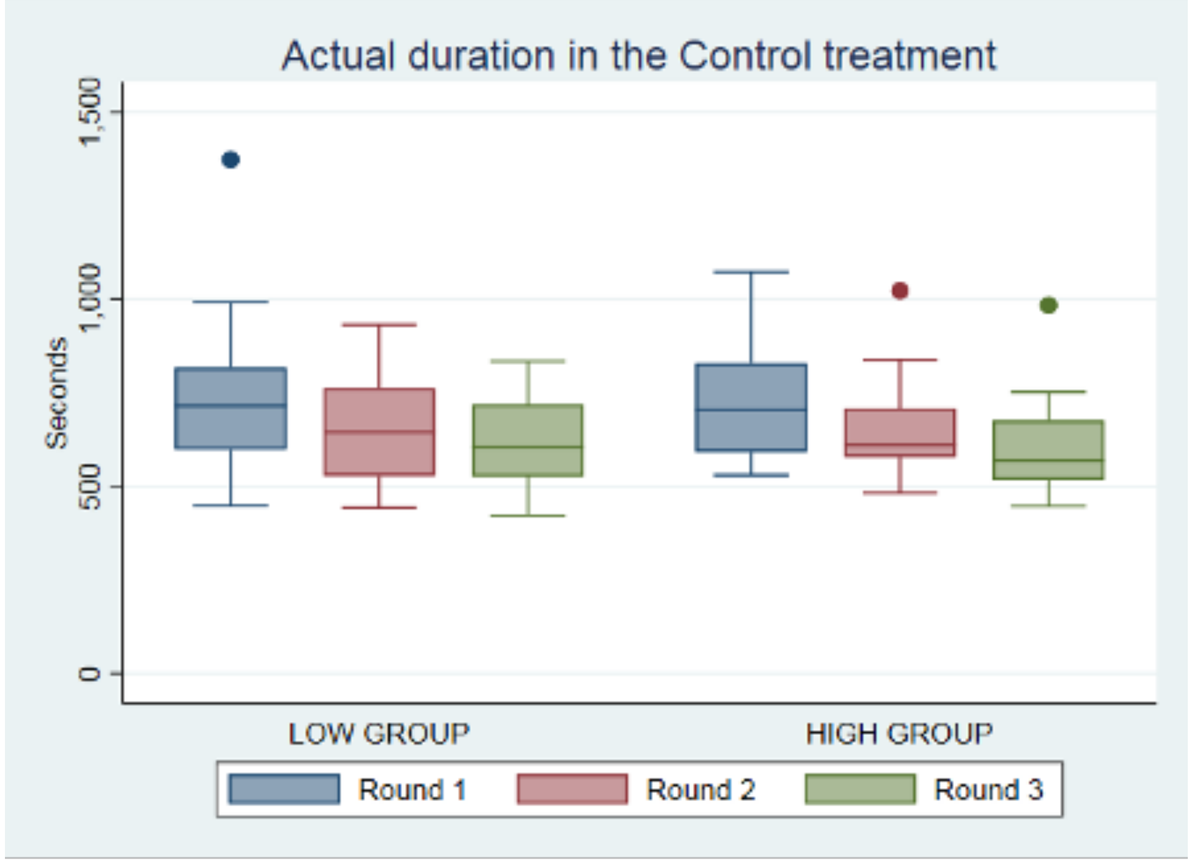


# Krabicový graf

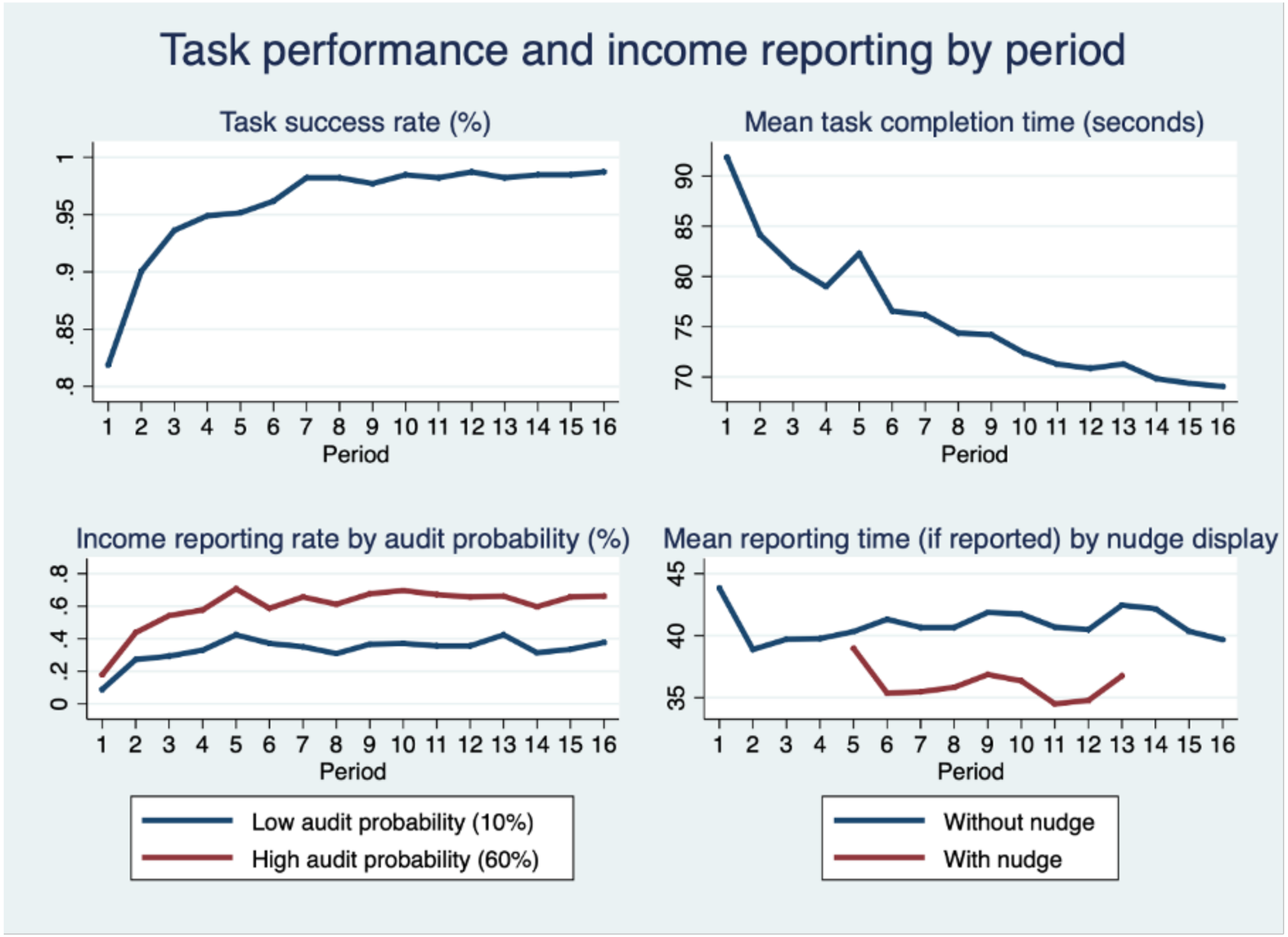
- Krabicový graf (boxplot) patrí tiež ku grafom, ktorými zobrazujeme kvantitatívne dáta. Je zložený z piatich základných hodnôt:
  - minimum (najnižšia hodnota premennej)
  - dolný kvartil (Q1, 25. percentil)
  - medián
  - horný kvartil (Q3, 75. per- centil)
  - maximum (najvyššia hodnota premennej).
- Inak povedané, v centre tohto grafu je medián, ktorý je obklopený zhora i zdola „krabicou“ a priestor oboch krabíc tvorí 50 % pozorovaní (medzikvartilový rozsah).
- Fúziky smerujúce nahor i nadol z krabíc označujú skóre Q3 + 1,5 medzikvartilového rozsahu (IQR) a Q1 + 1,5 IQR. Body nad/pod nimi predstavujú tzv. mimoležiace hodnoty, teda hodnoty nachádzajúce sa ďalej ako 1,5 IQR od hodnoty ohraničujúcej Q3.

# Krabicový graf

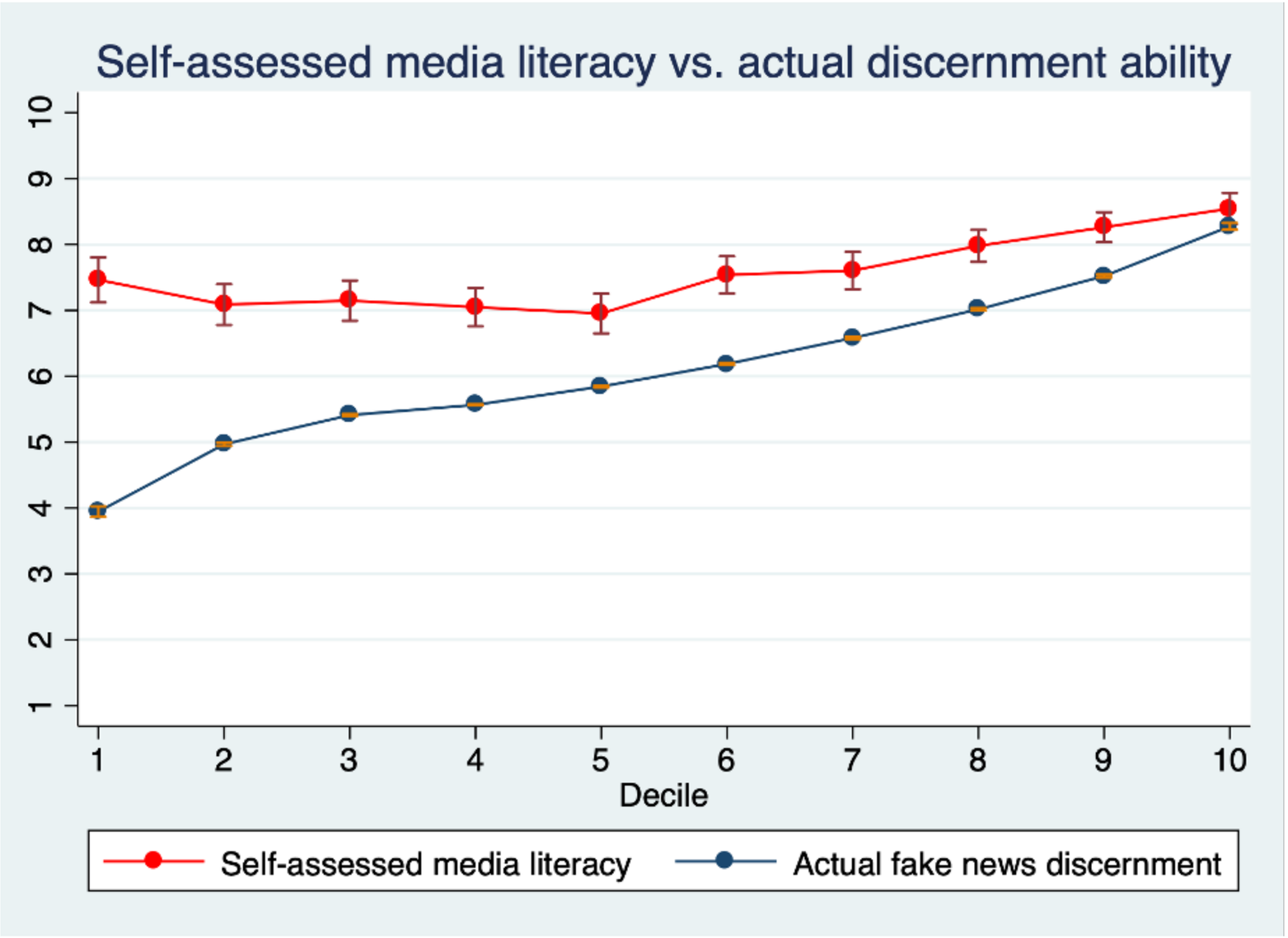


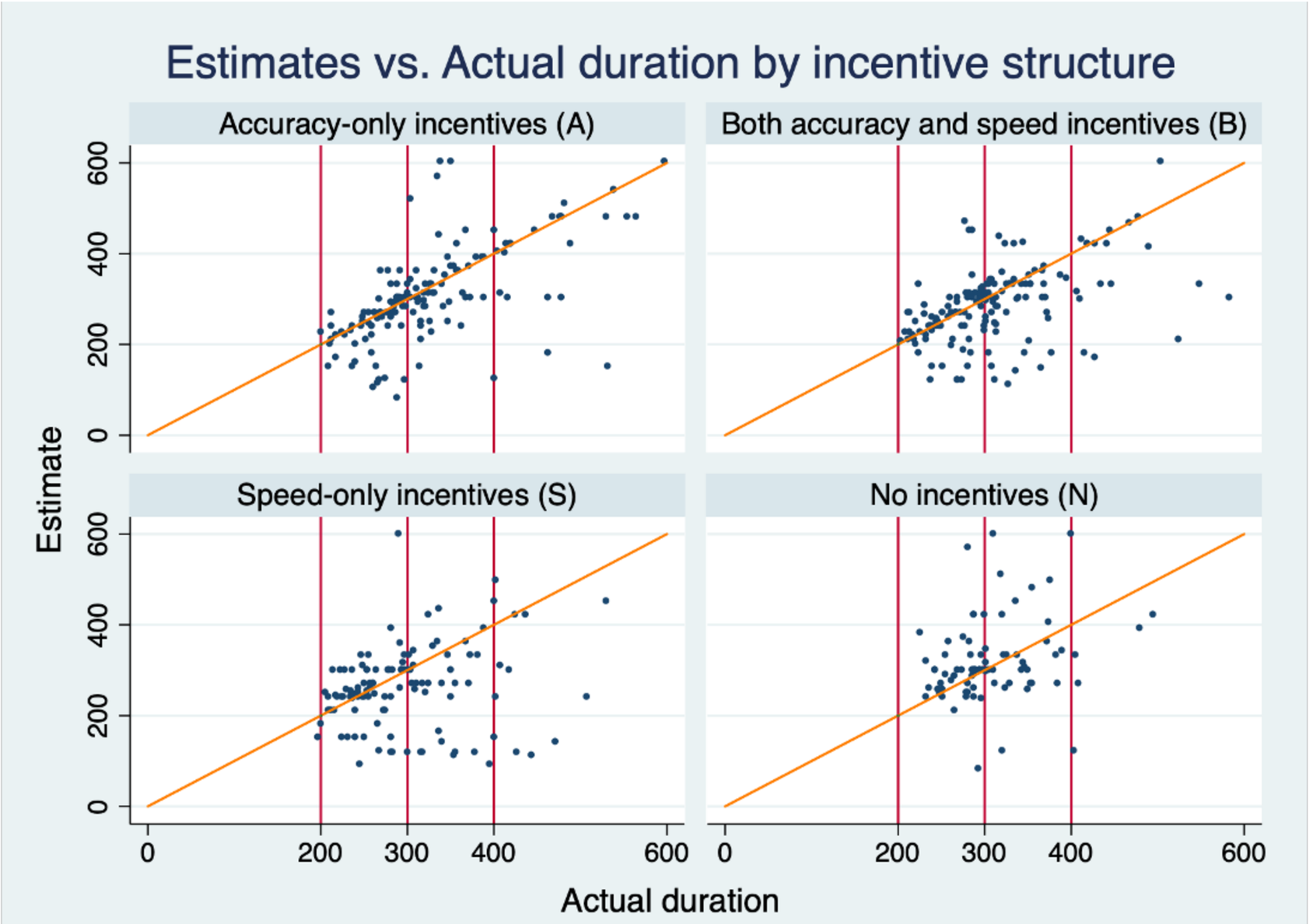
# Spojnicový graf



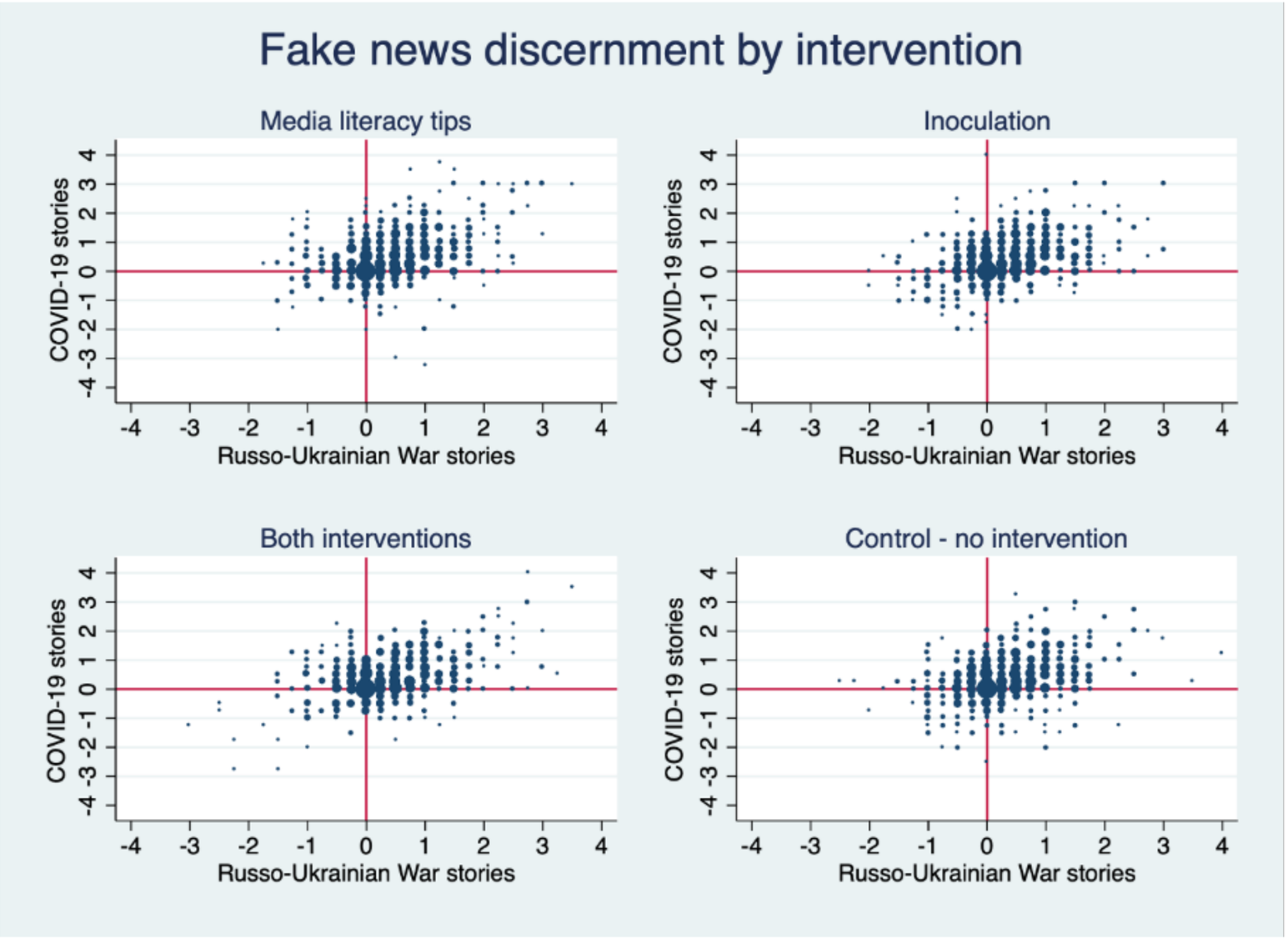
# Spojnicový graf



# Bodový diagram



# Bodový diagram



# Inferenčná štatistika

- V rámci skúmania sveta, ktorý je okolo nás, môže byť naším cieľom skúmaný jav iba jednoducho opísať. Môžeme sa pýtať napr. na to, koľko percent z opýtaných boli muži? Aký mali minimálny, maximálny a priemerný vek? Niekedy ale budeme chcieť naše zistenia zovšeobecniť nad rámec nášho výskumného súboru a dát, ktoré máme k dispozícii. Ide o tzv. inferenčnú štatistiku. Pri inferenčnej štatistike tvoríme závery o vzťahoch alebo rozdieloch na populačnej úrovni na základe dát získaných v našom výskumnom súbore.

- Pri využití inferenčnej štatistiky si môžeme položiť veľa otázok, ktoré by sme chceli preskúmať, formulovať veľké množstvo rôznorodých hypotéz, ktoré by sme chceli preveriť, a pri tom využiť množstvo rôznych postupov a konkrétnych druhov analýz. Ako si to zjednodušiť? Môžeme sa pokúsiť vytvoriť nejaké všeobecné kategórie inferenčnej štatistiky – vzťahovú a rozdielovú štatistiku a zvoliť dominantný prístup štatistickej inferencie – testtovanie nulovej hypotézy Rozdielová a vzťahová štatistika predstavujú dve strany tej istej mince, toto rozdelenie nám ale môže uľahčiť uvažovanie o niektorých problémoch a priniesť do nášho uvažovania systém.

- Pri vzťahovej štatistike nás bude zaujímať to, či je medzi dvoma premennými na populačnej úrovni vzťah a to, aký je tento vzťah tesný. Môžeme napr. predpokladať, že čím je človek vyšší, tým väčšiu váhu bude mať. V rámci vzťahovej štatistiky sa teda zameriame na spoločné kolísanie – kovarianciu – nameraných hodnôt. Môžeme napr. realizovať prierezový výskum, kde účastníkom administrujeme dotazník merajúci päť čŕt osobnosti a ďalšie premenné. V kontexte modelu osobnostných čŕt veľkej päťky by sme napr. mohli predpokladať, že čím sú ľudia viac otvorení skúsenosti (čím vyššie skóre vykazujú v danej dimenzii modelu veľkej päťky), tým pozitívnejší postoj budú mať voči menej konzervatívnym postupom výučby štatistiky, alebo čím vyššie budú skórovať v rámci dimenzie negatívnej emocionality, tým menej subjektívnej životnej pohody budú prežívať pred zápočtom zo štatistiky (tzv. predpokladáme negatívny vzťah medzi negatívnou emocionalitou a subjektívnou pohodou).

- Naproti tomu pri rozdielovej štatistike nás bude zaujímať rozdiel medzi skupinami alebo podmienkami. Môžeme napr. predpokladať, že muži budú vyšší než ženy. Otázka je tu položená inak než v predošlom prípade, a to skôr v intenciách toho, či je medzi danými skupinami rozdiel, a toho, aký veľký je tento rozdiel. V kontexte návrhu intervencií zlepšujúcich mentálne zdravie študentiek a študentov napríklad môžeme predpokladať, že študentky a študenti štatistiky, ktorí podstúpia nácvik určitého druhu relaxácie (napr. Jakobsonová progresívna relaxácia), budú na tom z hľadiska úrovne aktuálne prežívanej úzkosti pred cvičením lepšie než tí, ktorí relaxačnú metódu cvičiť nebudú. Pripravíme jednoduchý experiment, kde záujemkyne a záujemcov náhodne a/alebo na základe pretestu, zaradíme do experimentálnej a kontrolnej skupiny (medziskupinový experiment). S experimentálnou skupinou budeme raz denne cvičiť relaxačnú metódu, zatiaľ čo s kontrolnou skupinou sa iba stretneme, relaxáciu však cvičiť nebudeme. Po mesiaci porovnáme obe skóre v subjektívne prežívanej úzkosti.

- Participantky a participantov však nemusíme nevyhnutne deliť do skupín. Namiesto toho sa nám môže hodiť, aby všetci absolvovali všetky pod- mienky, ale v rôznom poradí (tzv. vnútrosubjektový  experiment). V kontexte výskumu efektívnosti reklám napr. môžeme predpokladať, že obsah (inzero- vaný produkt) vtipnej reklamy na štatistickú knihu bude hodnotený pozitívnejšie, než pri neutrálnej reklame. Následne zobrazíme každej participant- ke/participantovi nášho výskumu v náhodnom poradí neutrálnu a vtipnú reklamu. Pri každej z nich označia to, ako pozitívne hodnotia štatistickú knihu zobrazenú na obrázku. Na záver porovnáme hodnotenia produktu tých istých participantiek/participantov v rámci vtipnej aj neutrálnej podmienky.

# Pearsonova korelácia

- Koralácia meria stupeň asociácie medzi dvoma premennými. Základným nnástrojom je tzv. Pearsonova korelácia, ktorá meria stupeň linearity dvojrozmerného vzťahu medzi dvoma premennými.

- Pearsonova korelácia je štandardizovaná kovariancia (t. j. pomer kovariancie medzi dvoma premennými k súčinu ich rozptylov). Pearsonova korelácia je ohraničená medzi -1 (dokonalý negatívny lineárny vzťah) a 1 (dokonalý pozitívny lineárny vzťah).

- Pearsonovu koreláciu neovplyvňujú zmeny v rozsahu (napr. vynásobenie premenných 2) ani v umiestnení (napr. pridanie konštanty) premenných. Dve úplne nezávislé premenné majú korelačný koeficient rovný 0. Zistenie nulovej korelácie medzi dvoma premennými však možno interpretovať ako indikáciu nezávislosti len v špeciálnych prípadoch, napr. pri dvojrozmernom normálnom rozdelení. Použitie Pearsonovej korelácie by tu bolo zavádzajúce vzhľadom na jej základný predpoklad linearity.

- Obrázok na ďalšom slajde predstavuje Anscombovo kvarteto, sériu štyroch súborov údajov, ktoré zostavil Anscombe (1973), aby zdôraznil význam vizualizácie údajov pred vykonaním štatistickej analýzy. Každý súbor údajov pozostáva z 11 dvojíc bodov s Pearsonovou koreláciou 0,816. Priemer prvej premennej (na osi x) je vždy 9, pričom výberový rozptyl je 11. Priemer druhej premennej je približne 7,50, pričom výberový rozptyl je od 4,122 do 4,127. Pearsonova korelácia má zmysel len v paneli a), pre normálne rozdelené údaje s lineárnym vzťahom. V paneli b) je vzťah jednoznačne nelineárny. Panely c) a d) ukazujú, ako je Pearsonova korelácia citlivá na odľahlé hodnoty. V paneli c) je bez odľahlých hodnôt korelácia 1. Stačí jedna jediná odľahlá hodnota, aby sa znížila na 0,816. V paneli d) je korelácia bez odľahlej hodnoty 0, ale stačí jedna odľahlá hodnota, aby sa zvýšila na 0,816.

# Pearsonova korelácia