TNG TECHNOLOGY CONSULTING

The statistics of effort

$\pi \times$ 

Dr. Steffen Rath, TNG

Big Techday 11, May 18th, 2018, Munich

TNG TECHNOLOGY CONSULTING

Who are you?

- Are you among those who…
  - … Receive estimations
  - … Give estimations
- Do you perceive estimations as…
  - … Helpful?
  - … An instrument of torture?
- Do you feel you know how far you can trust your estimations?

TNG TECHNOLOGY CONSULTING

A practical exercise

Let‘s do an estimation! Please estimate the following:

Read and understand

- a short text (~35 words)
- in German

Example text 1

Hänschen klein
Ging allein
In die weite Welt hinein.
Stock und Hut
Steht ihm gut,
Ist gar wohlgemut.
Aber Mutter weinet sehr,
Hat ja nun kein Hänschen mehr!
"Wünsch dir Glück!"
Sagt ihr Blick,
"Kehr' nur bald zurück!"

Franz Wiedemann (1821-1882)

Example text 2

Die Einheit der horizontalen Schemata von Zukunft, Gewesenheit und Gegenwart gründet in der ekstatischen Einheit der Zeitlichkeit.

Der Horizont der ganzen Zeitlichkeit bestimmt das, woraufhin das faktisch existierende Seiende wesenhaft erschlossen ist.

Martin Heidegger (1889-1976)

TNG TECHNOLOGY CONSULTING

Findings from exercise

- Different durations for the two texts
- Differences pretty obvious a posteriori
- Two strategies to deal with uncertainty
  - Reduce the lack of information
    - Can make estimates more precise
    - Requires additional effort for each estimation
    - No plan for estimations that are still far off
  - Understand it and plan for it
    - One-time effort for understanding
    - Occasional effort to do statistics
    - Errors are part of the game

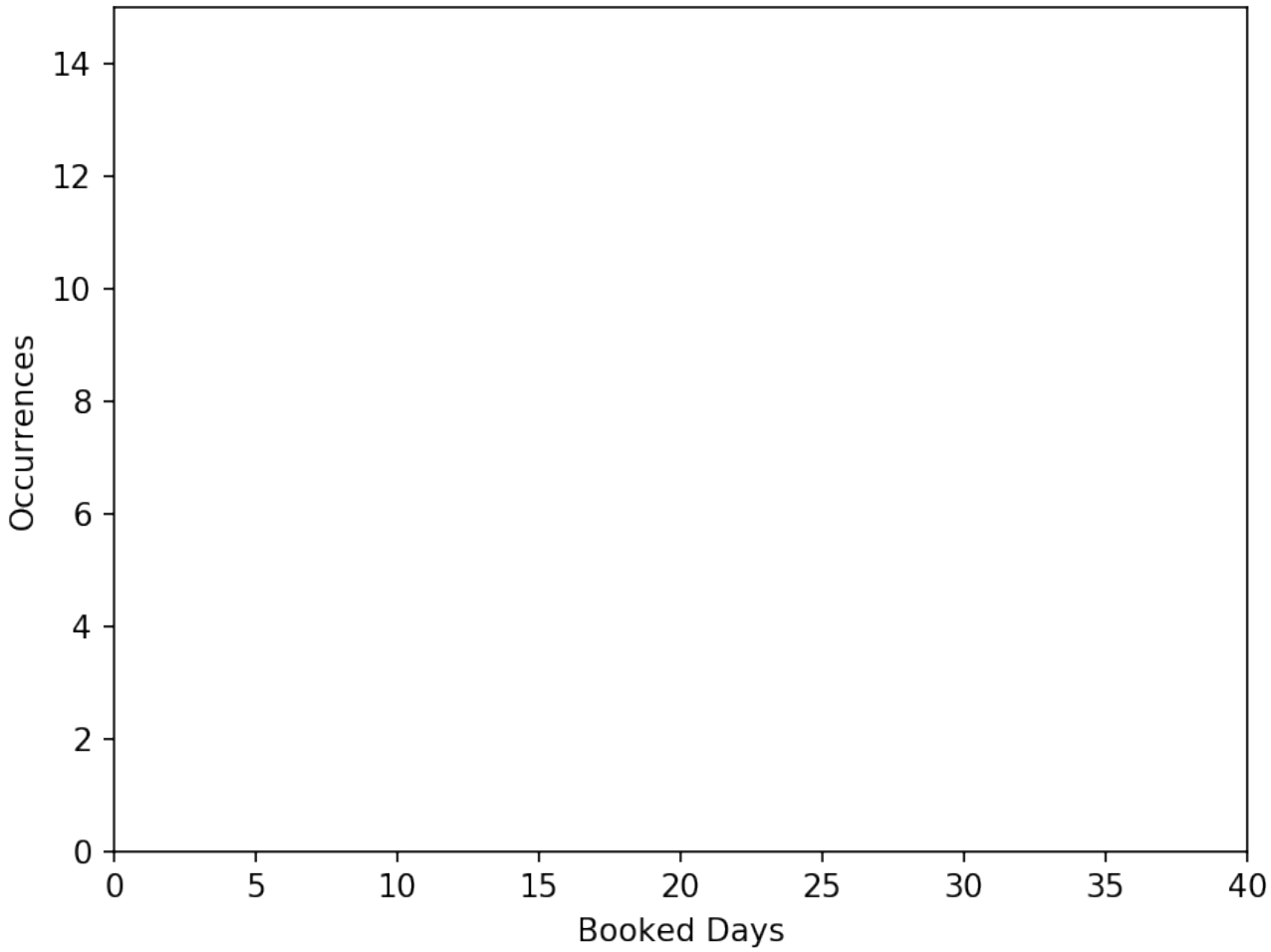
TNG TECHNOLOGY CONSULTING

Estimations in „real life“

- Imagine a big feature request – what will it cost?
  - Upper bound will not do due to budgeting constraints
  - NB: „When will it be done“ is a different question!
- Get estimates
  - Break down into smaller bits for development
  - Have developers estimate effort for each part
  - Total estimate is sum of estimates of parts
- Greenlight development
- Total effort regularly larger than estimated
- What did we do wrong?

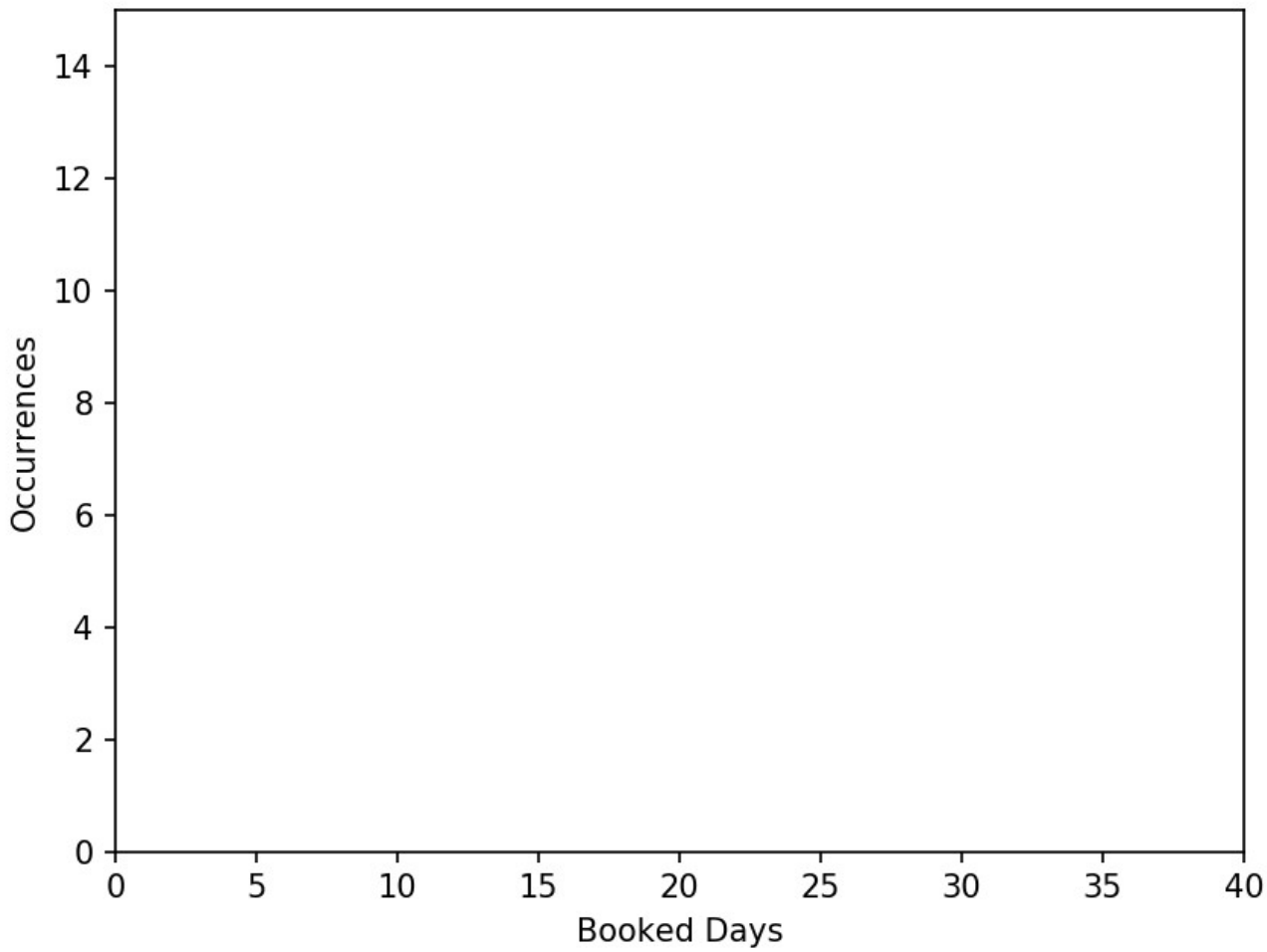
TNG TECHNOLOGY CONSULTING

A typical histogram



TNG TECHNOLOGY CONSULTING

A typical histogram



TNG TECHNOLOGY CONSULTING

Properties of the histogram

- Uni-modal with clear maximum
- Quite broad
- Asymmetric
- Many outliers
- Any calculation plagued by noise
- With more data points (hypothetical):
  - Gets smoother
  - Allows more precise calculations
  - Approaches (scaled) probability distribution

TNG TECHNOLOGY CONSULTING

Outline

Some basics on statistics

The precision of estimations

Checking for consistency

TNG TECHNOLOGY CONSULTING

Anatomy of a probability distribution

- Normalized to unity, e.g.,

$$\sum_{n \in \mathbb{Z}} p(n) = 1$$

- Maximum (if present): most probable value
  - Best answer to question: „How long will it take you?“
  - Usually not the question we want to answer

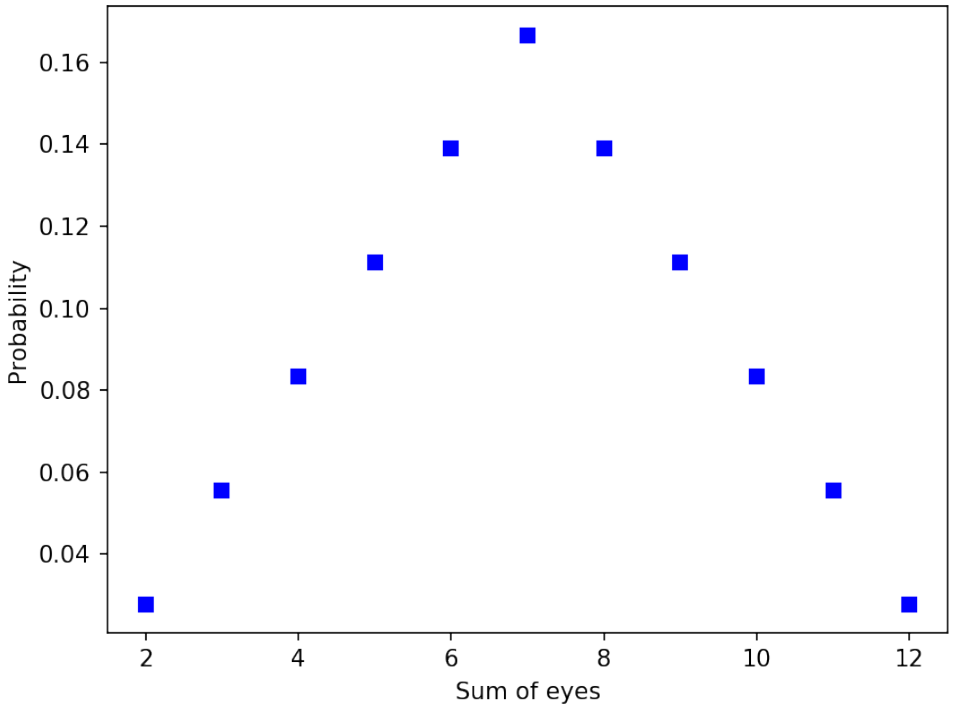
- Familiar example: probability distribution for an ideal die

$$p(n) = \begin{cases} 1/6 & n \in \{1, \ldots, 6\} \\ 0 & \text{otherwise} \end{cases}$$

Probability distribution for sums

- Consider throwing two dice. Probability for throwing a total of 10:

$$(p * p)(10) = p(4)p(6) + p(5)p(5) + p(6)p(4) = 1/12$$

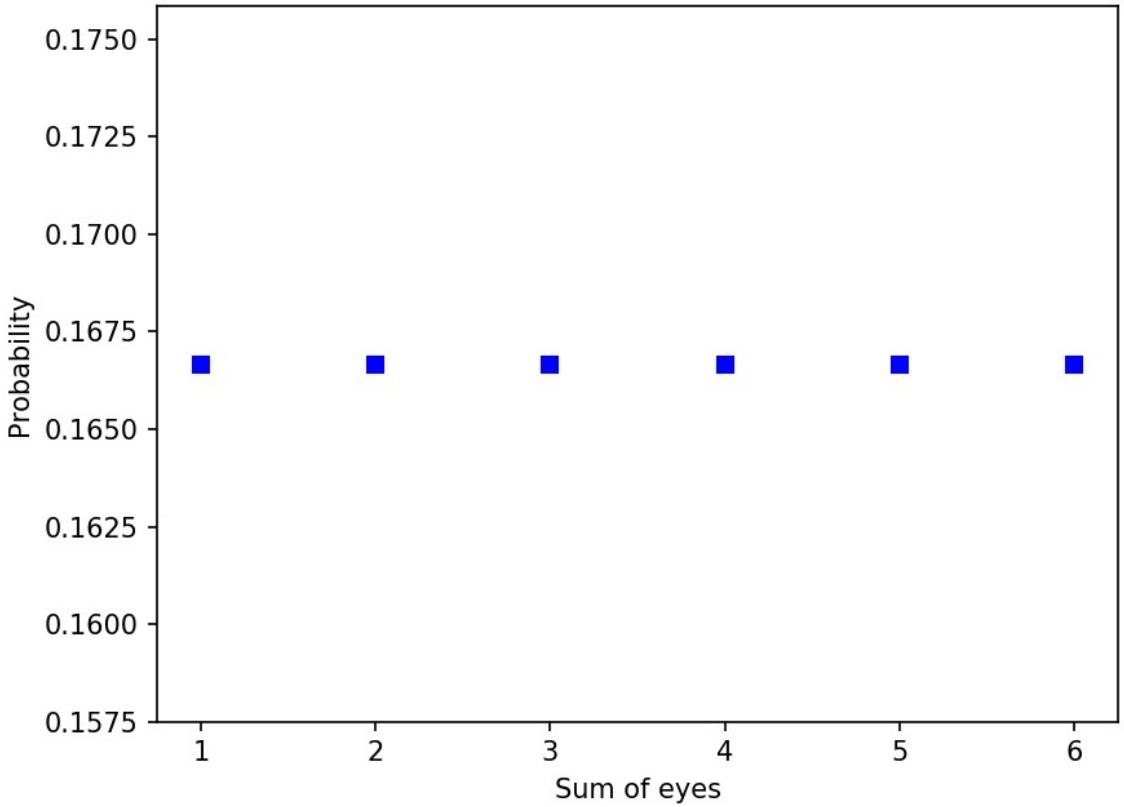


- Probability distribution for a sum looks quite different

TNG TECHNOLOGY CONSULTING

Adding up more things

- How does the distribution develop with the number of dice?



TNG TECHNOLOGY CONSULTING

Primer on statistics basics: moments

Definition of an expectation with respect to distribution *p*:

$$\langle f(n) \rangle_p = \sum_n f(n) p(n)$$

Special case: **moments**

$$f(n) = n^k \qquad \Rightarrow \qquad \langle n^k \rangle_p = \sum_n n^k p(n)$$

First moment („**the“ expectation**) has a very nice property:

$$\langle n \rangle_{p*q} = \langle n \rangle_p + \langle n \rangle_q$$

None of the other moments has this property.

TNG TECHNOLOGY CONSULTING

Primer on statistics basics: cumulants

Polynomials of the moments with the property of additivity:

$$\langle n^k \rangle_{p*q,c} = \langle n^k \rangle_{p,c} + \langle n^k \rangle_{q,c}$$

First three cumulants:

$$\begin{aligned}
\langle n \rangle_c &= \langle n \rangle \\
\langle n^2 \rangle_c &= \langle n^2 \rangle - \langle n \rangle^2 \\
\langle n^3 \rangle_c &= \langle n^3 \rangle - 3\langle n^2 \rangle \langle n \rangle + 2\langle n \rangle^3
\end{aligned}$$

Second cumulant: **variance** (squared **standard deviation**)

$$\left\langle (n - \langle n \rangle)^2 \right\rangle$$

Standard deviation: characterises the **width** of a distribution

TNG TECHNOLOGY CONSULTING

Primer on statistics basics: cumulants

Scaling when adding lots of random variables:

$$\langle x^k \rangle_c \mapsto N \langle x^k \rangle_c \quad \Rightarrow \quad \frac{\langle x^k \rangle_c}{\langle x \rangle^k} \mapsto \frac{1}{N^{k-1}} \frac{\langle x^k \rangle_c}{\langle x \rangle^k}$$

- Higher cumulants increasingly negligible as *N* increases
- Standard deviation decreases like $1/\sqrt{N}$

Limit as *N* tends to infinity: **normal distribution** (Gaussian)

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

Primer on statistics bas

Scaling beha

The estimation fallacy

Our distributions are asymmetric

- Expected: effort bounded from below, but not above
- Consequence: **expectation** > **most probable value**

Estimation using abstraction (e.g., story points)

- Distinguishes only between „complexity classes“
- Agnostic to expectation and most probable value

Estimation using man days

- Estimate single story: try to hit most probable value
- Many stories: average tends to expectation
- Consequence: estimations without abstraction systematically underestimate the effort

TNG TECHNOLOGY CONSULTING

Outline

Some basics on statistics

The precision of estimations

Checking for consistency

TNG TECHNOLOGY CONSULTING

Data used for statistics

Effort tracking in medium-sized software development project

- User Stories estimated in abstract story points
- Data spans two years with four releases / year
- Used data fields:
  - Story points
  - Total booked time (including story and all subtasks)
  - Resolution date

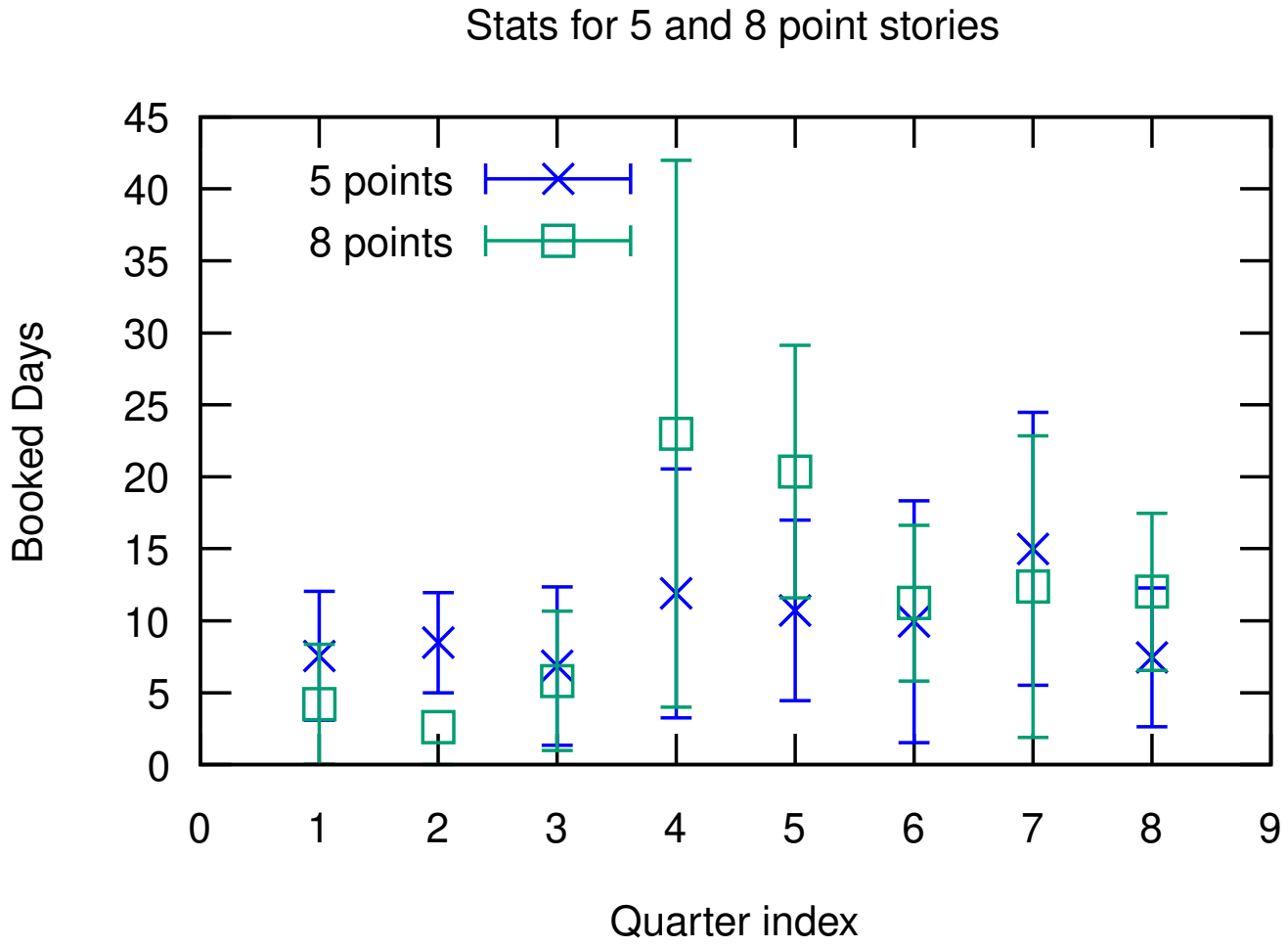
Disclaimer: We use terms from agile processes for brevity, the project was mostly waterfall

Story points	1	2	3	5	8	13	total
# Stories	26	80	127	134	56	26	449

TNG TECHNOLOGY CONSULTING

Expectation and standard deviation

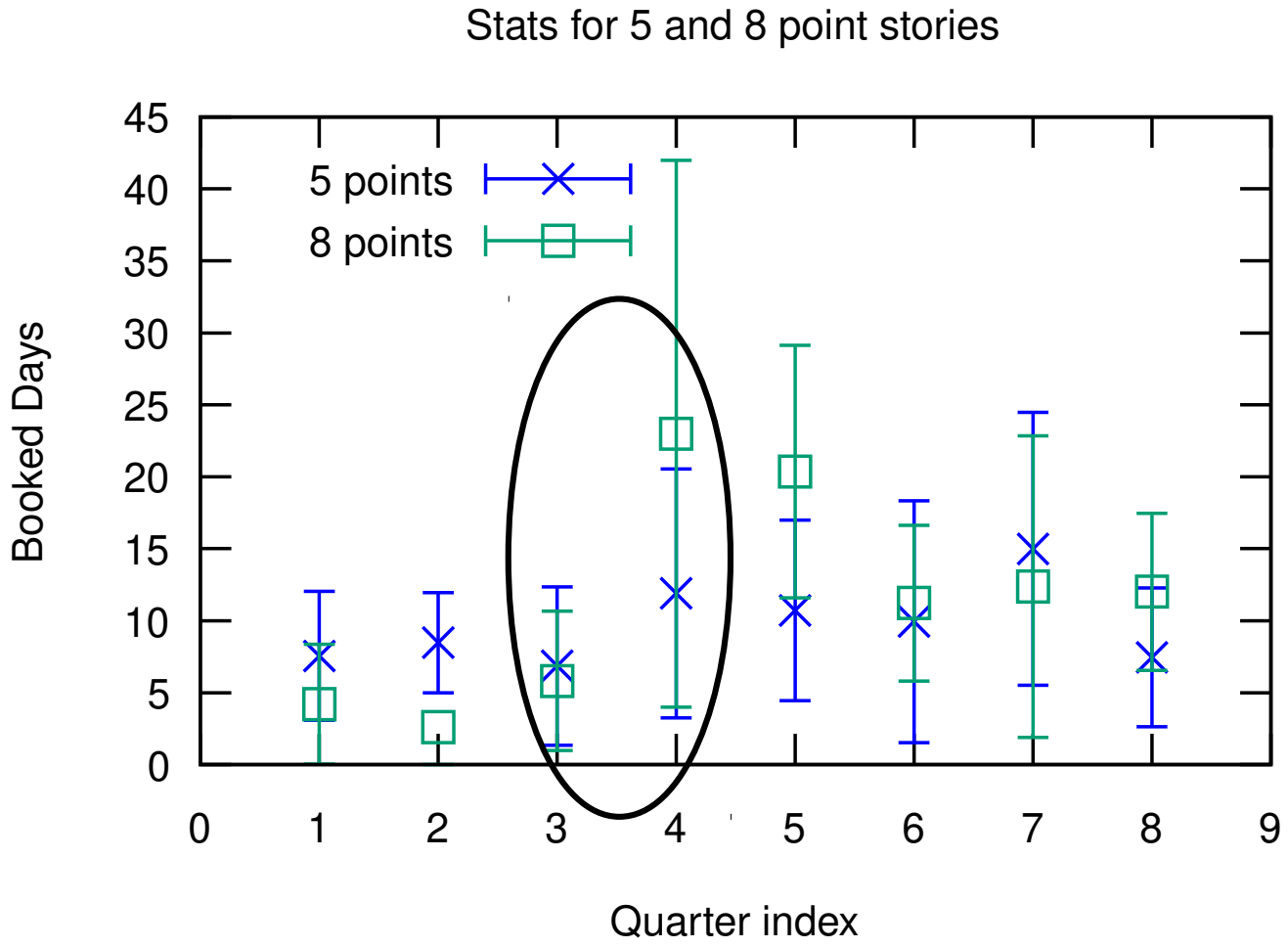
Average and standard deviation quarter by quarter:



TNG TECHNOLOGY CONSULTING

Expectation and standard deviation

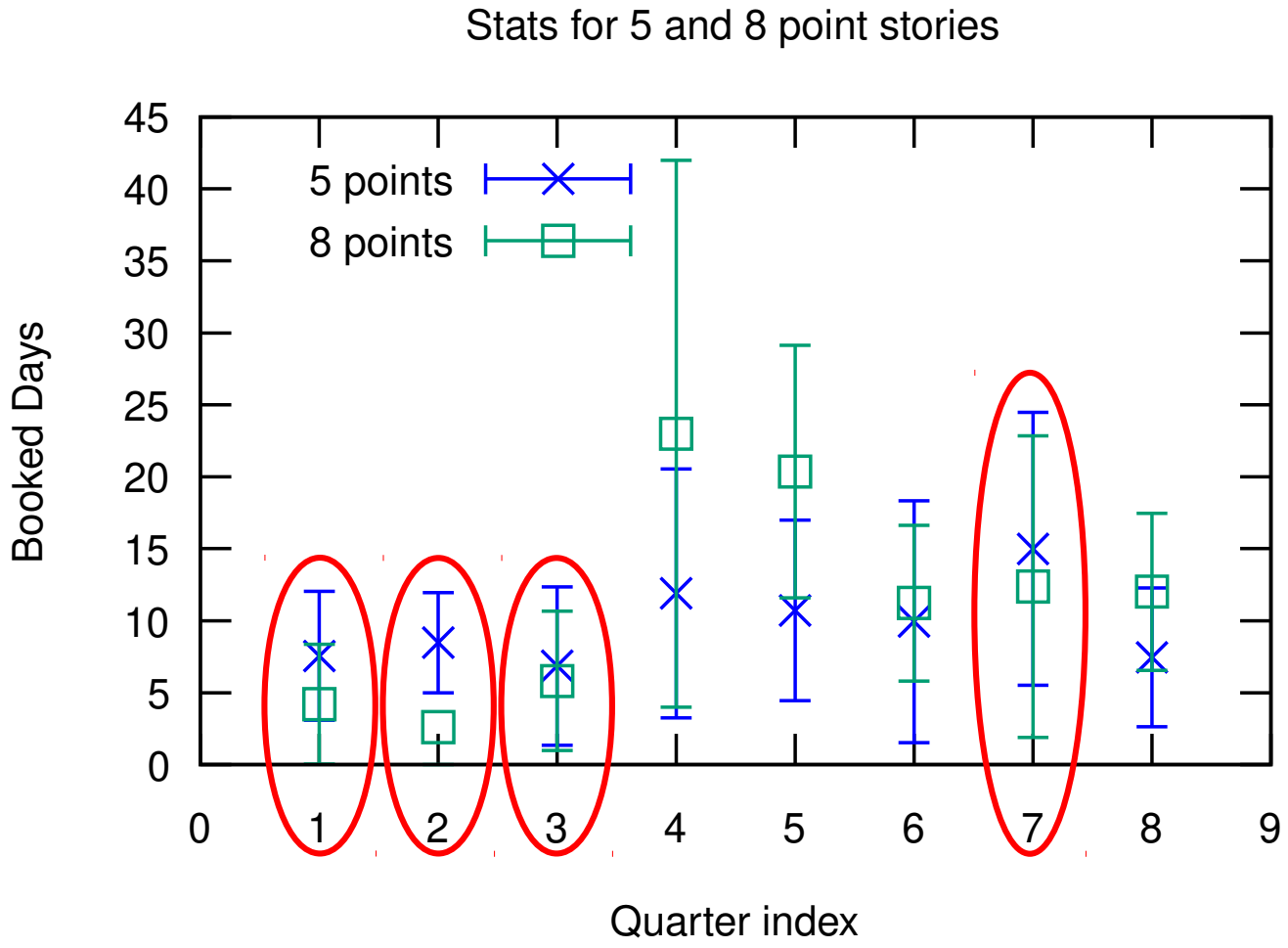
Average and standard deviation quarter by quarter:



TNG TECHNOLOGY CONSULTING

Expectation and standard deviation

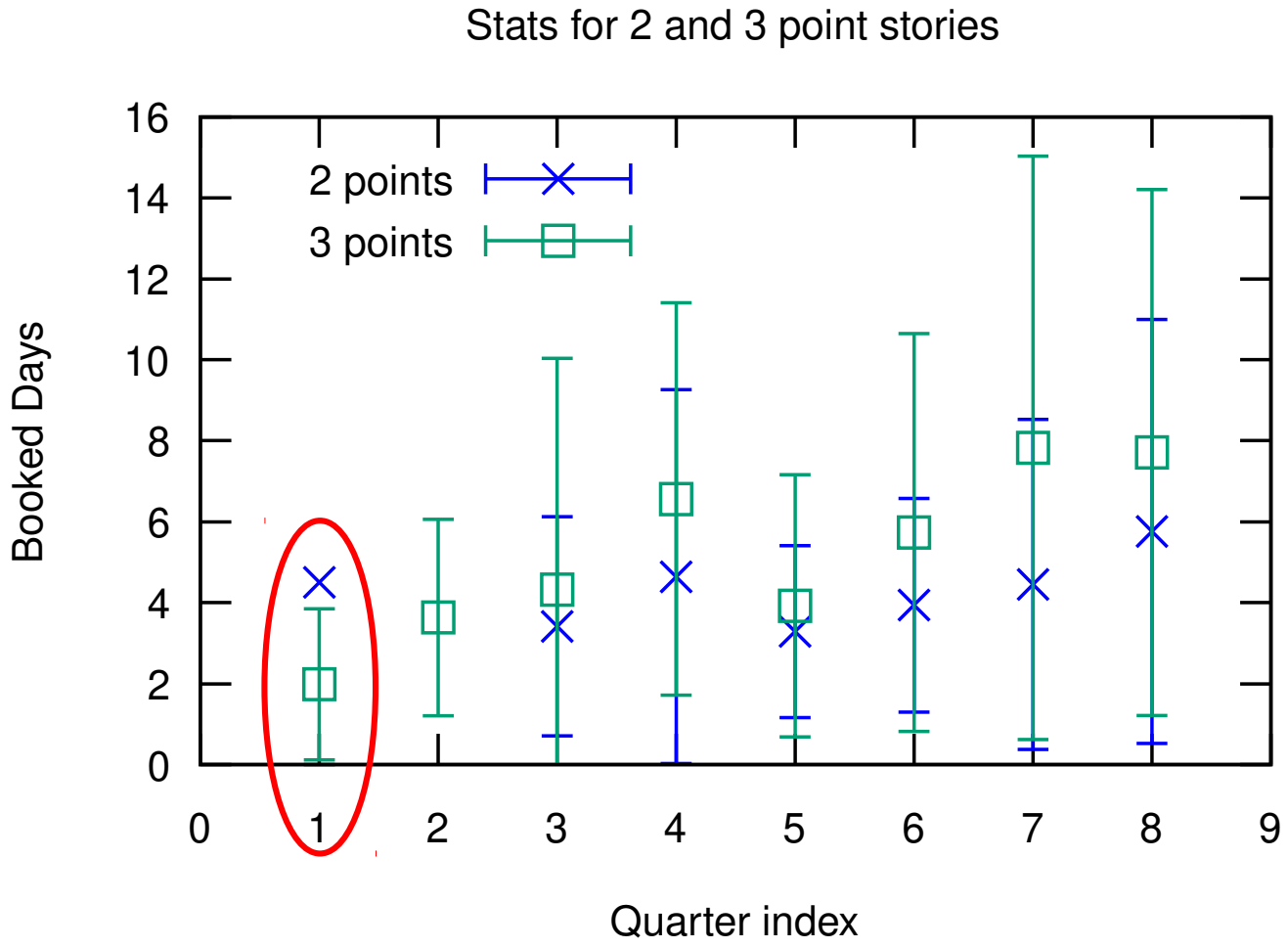
Average and standard deviation quarter by quarter:



TNG TECHNOLOGY CONSULTING

Expectation and standard deviation

5 and 8 points is the worst, here something more typical:



TNG TECHNOLOGY CONSULTING

What do we mean by 10% accuracy?

- Sufficiently many stories: sum of complexities may be assumed to have normal distribution.
- Fraction of values within 1, 2, 3 std deviations: 68%, 95% and 99.7%
- Assuming relative width per story of 0.7 (best case), one needs 49 stories for the sum to be within 10% with 68% security:
  - Relative width needs to decrease by a factor of 7
  - $\sqrt{N} = 7 \Rightarrow N = 49$
- For 95% security, we need 4 * 49 = 196 stories
  - Because 10% now is to **two** standard deviations
- For 99.7% security, we need 9 * 49 = 461 stories

TNG TECHNOLOGY CONSULTING

Outline

Some basics on statistics

The precision of estimations

Checking for consistency

TNG TECHNOLOGY CONSULTING

A quantitative test of consistency

Hard to tell whether data is consistent over time

- Relatively small samples
- Very broad distributions
- Large fluctuations of expectations

The Mann-Whitney *U* test

- Helps decide whether two samples are drawn from same distribution
- No assumption about the functional form of distribution
- Works as **significance test**

TNG TECHNOLOGY CONSULTING

Mann-Whitney U test: principle

Input: two sets of numbers:

$$M = \{3, 10, 12, 17, 19\} \quad ; \quad N = \{4, 8, 13, 14, 18\}$$

Determine the **ranks** of these numbers:

3		10	12			17		19	
	4	8			13	14		18	
1	2	3	4	5	6	7	8	9	10

Rank sums:

$$R_M = 1 + 4 + 5 + 8 + 10 = 28 \quad ; \quad R_N = 2 + 3 + 6 + 7 + 9 = 27$$

Definition of *U* (equivalently for *N*):

$$U_M = R_M - \frac{|M|(|M| + 1)}{2} = 28 - 15 = 13$$

U is normally distributed with known expectation and variance

TNG TECHNOLOGY CONSULTING

Mann-Whitney U test: significance testing

- Define significance threshold: $p_c = 0.05$
  - 5% chance of **rejecting** the null hypothesis erroneously
  - does **not** mean 5% chance of **retaining** it erroneously

Numbers from previous slide:

$$U = 13 \quad ; \quad \mu_U = 12.5 \quad ; \quad \sigma_U^2 = 22.92\ldots$$

Probability that normal random value differs from expectation by more than $U - \mu$

$$p = 2 \int_{|U-\mu|}^{\infty} dx \frac{\exp(-x^2/2\sigma^2)}{\sqrt{2\pi\sigma^2}}$$

Here $p = 0.917$, so we retain the null hypothesis

TNG TECHNOLOGY CONSULTING

Temporal consistency

- For sufficient sample size, used half-year „time slices“
- Applied *U* test to all pairs of time slices for stories with same complexity

Examples from the result set:

5 point stories

Time slice index	1	2	3	4
1	1	0.890	0.431	0.645
2		1	0.474	0.688
3			1	0.653
4				1

8 point stories

Time slice index	1	2	3	4
1	1	—	—	—
2		1	0.046	0.199
3			1	0.377
4				1

- Very few values below threshold
- The values below threshold are correlated with the jump in the graphs
- Fluctuations in agreement with consistent estimations

TNG TECHNOLOGY CONSULTING

Precision of the used scale

- Can we claim distributions for adjacent numbers of points are actually different?
- Results from entire data
- NB: here values below threshold are a **good** sign!

Story points	1	2	3	5	8	13
1	1	0.143	0.023			
2		1	0.234	6.83×10^{-10}		
3			1	1.77×10^{-8}	4.20×10^{-5}	
5				1	0.559	5.40×10^{-5}
8					1	0.0029
13						1

TNG TECHNOLOGY CONSULTING

Precision of the used scale

- Same thing considering only stories after the jump

Story points	1	2	3	5	8	13
1	1	0.140	0.0056			
2		1	0.044	2.26×10^{-9}		
3			1	3.91×10^{-6}	4.49×10^{-7}	
5				1	0.020	1.83×10^{-7}
8					1	0.00029
13						1

TNG TECHNOLOGY CONSULTING

Coarse graining the scale

- Replace story point scale by „T shirt sizes“:
  - Small: 1-3 points
  - Medium: 5-8 points
  - Large: 13+ points
- Relative widths between 0.8 and 0.95 (vs 0.7 to 0.9)
- Time slice consistency slightly better than with story points
- Yields clearly different distributions (using all data):

Size	small	medium	large
small	1	3.93×10^{-16}	9.24×10^{-11}
medium		1	8.83×10^{-5}
large			1

TNG TECHNOLOGY CONSULTING

Conclusion

Results of the analysis

- Estimations are imprecise – we have to deal with it
- Distributions are asymmetric
  - Beware of difference between most probable value and expectation
  - Abstractions such as story points / T shirt sizes are helpful
- As long as conditions do not change too rapidly, estimations are consistent over time
- The precision of the story point scale is just barely justifiable* – T shirt sizes would be a better fit

*** for the investigated project**