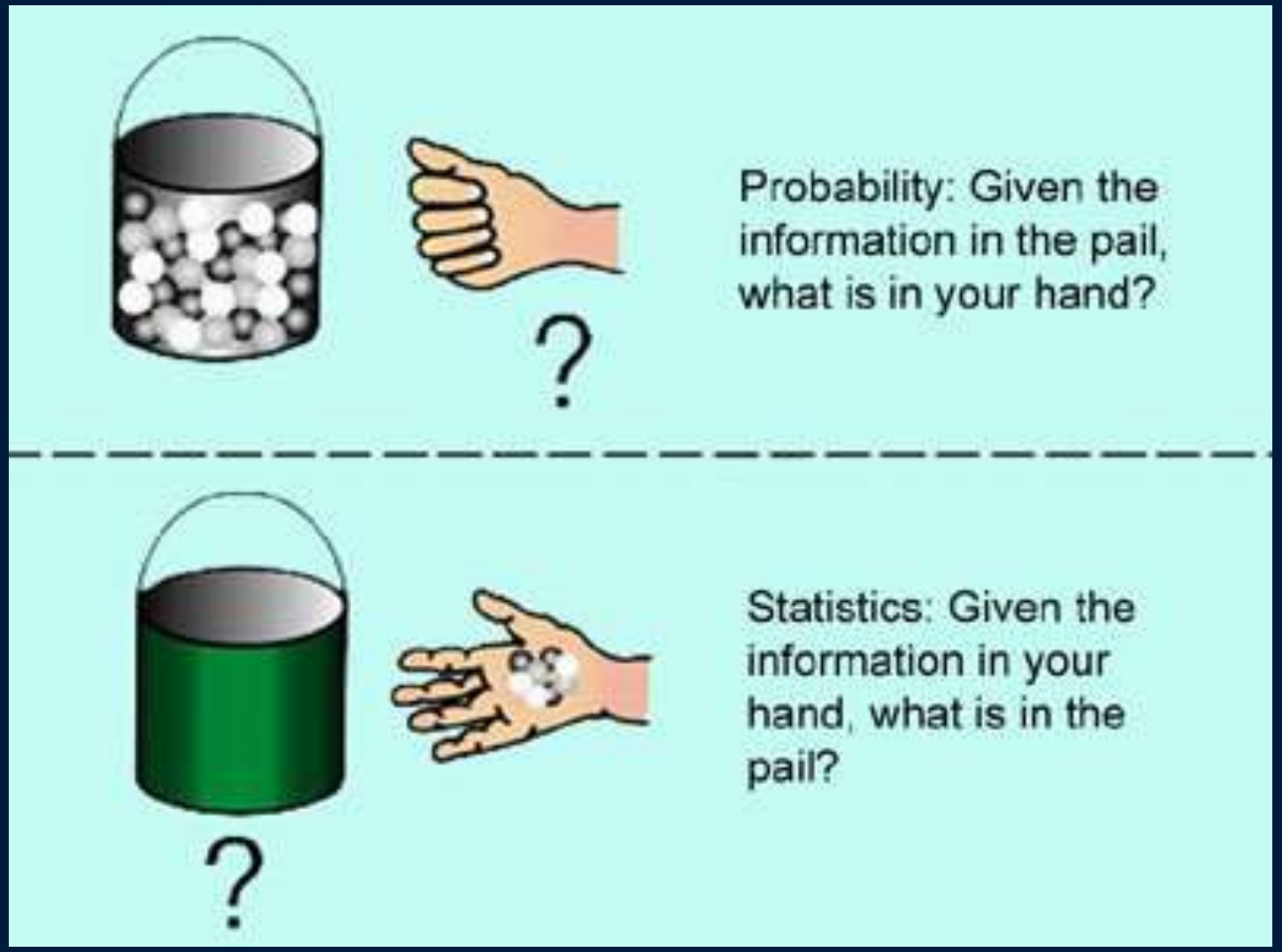
BCS1520: Statistics

Lecture 06:
Probability meets Statistics

Maastricht University | Department of Advanced Computing Sciences

Probability Theory vs Statistics

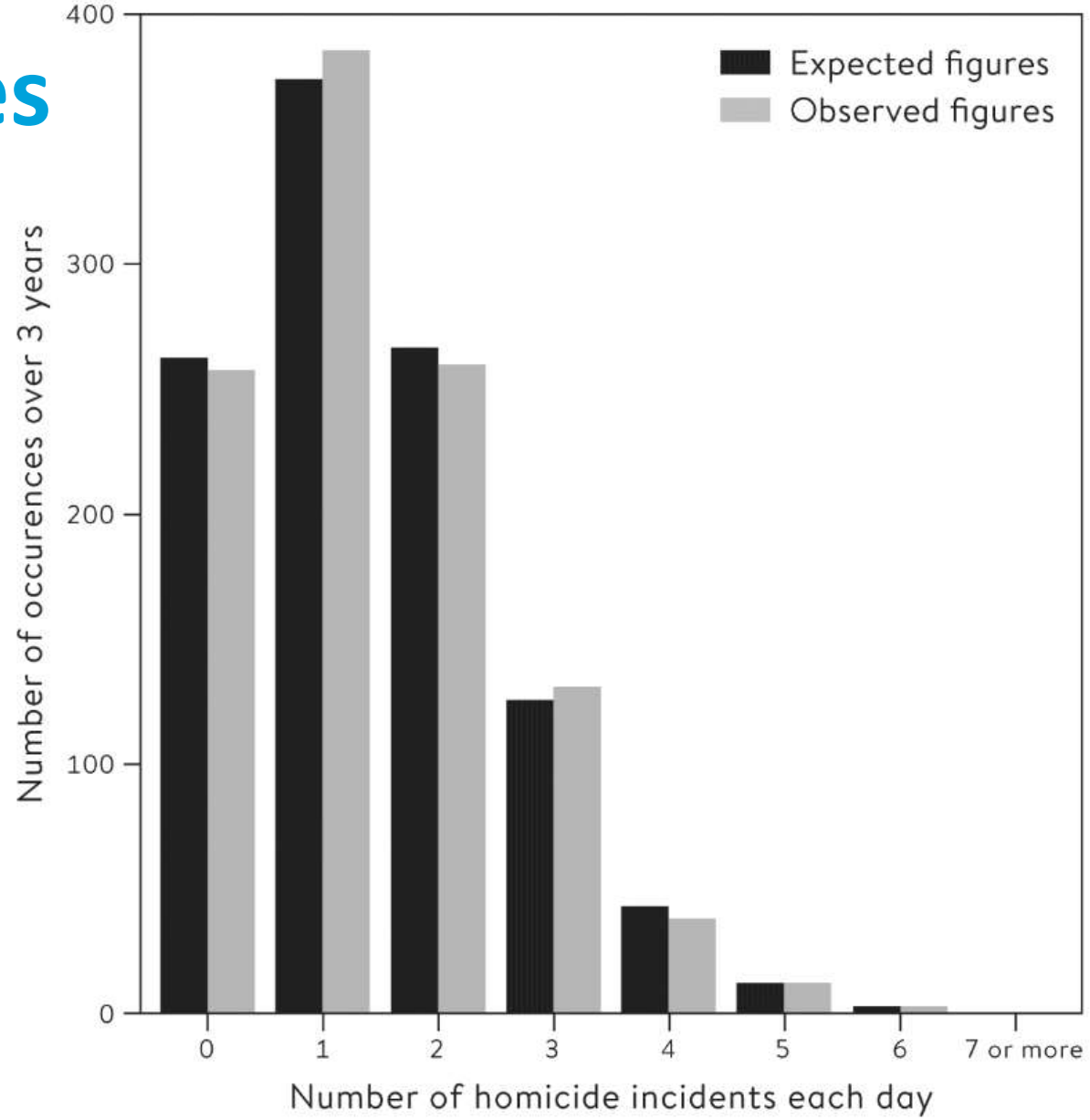


Maastricht University

Picture from: https://www.quora.com/What-is-the-difference-between-probability-and-statistics

Discrete Random Variables

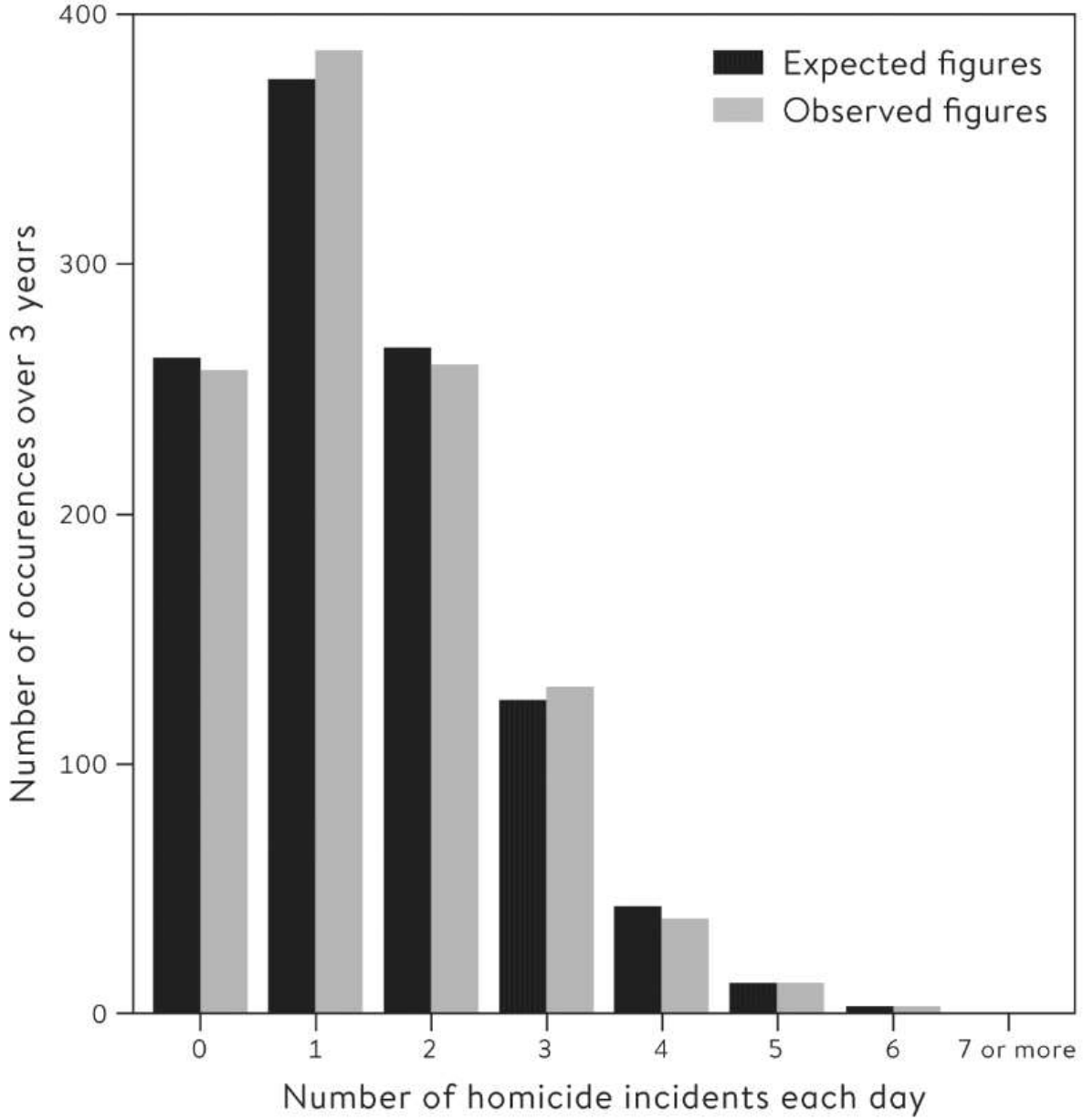
- Thinking about the Poisson distribution for homicide events again.
- Every day is considered an ***independent*** random variable.
- So for each day, we are drawing from the Poisson distribution.



Independent RVs

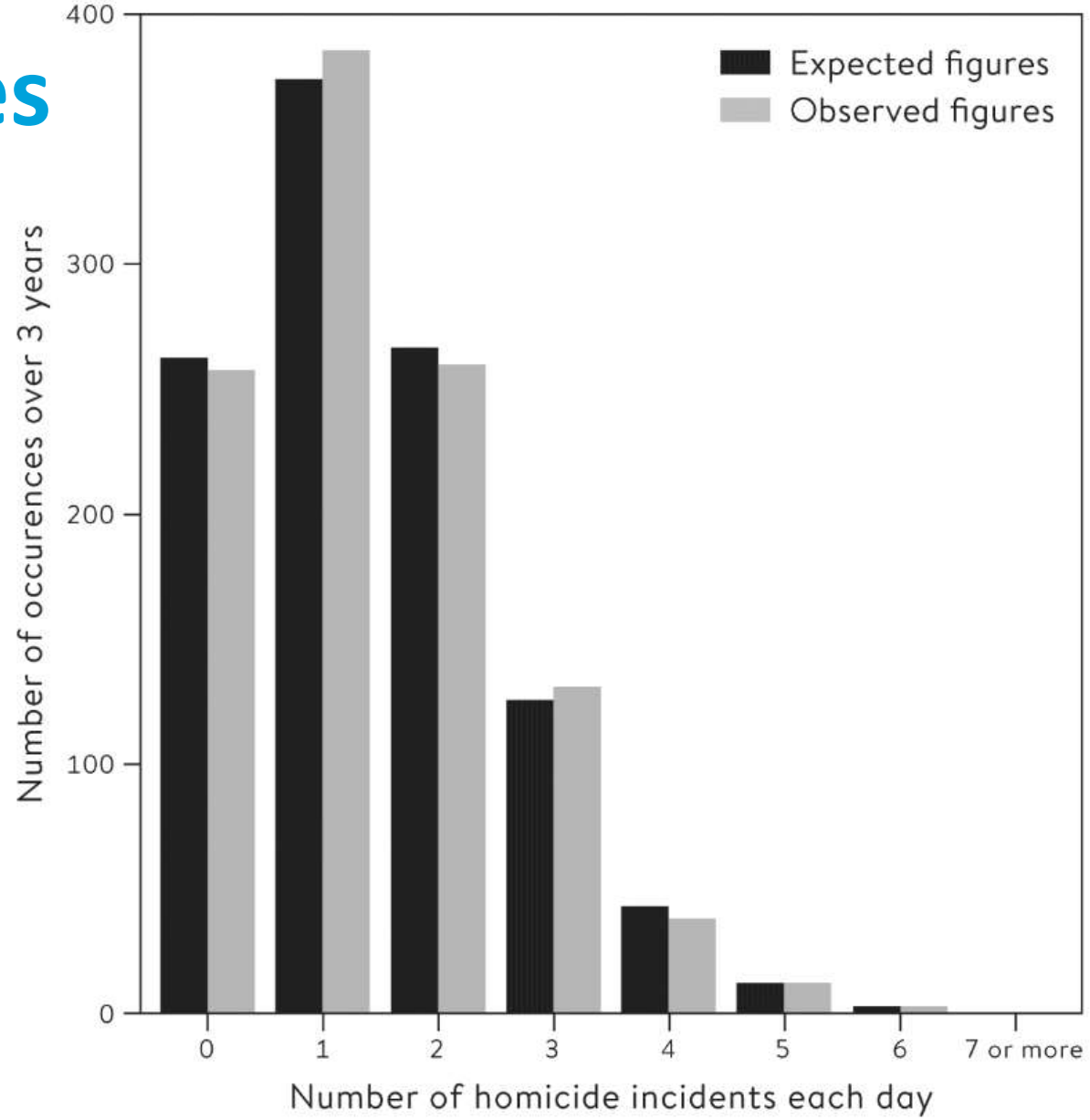
Every day is considered an ***independent*** random variable.

- Consistently an assumption when using probability theory to do statistics.
- We assume each sample we have is an independent random variable.
- OR. We model the non-independence directly.



Discrete Random Variables

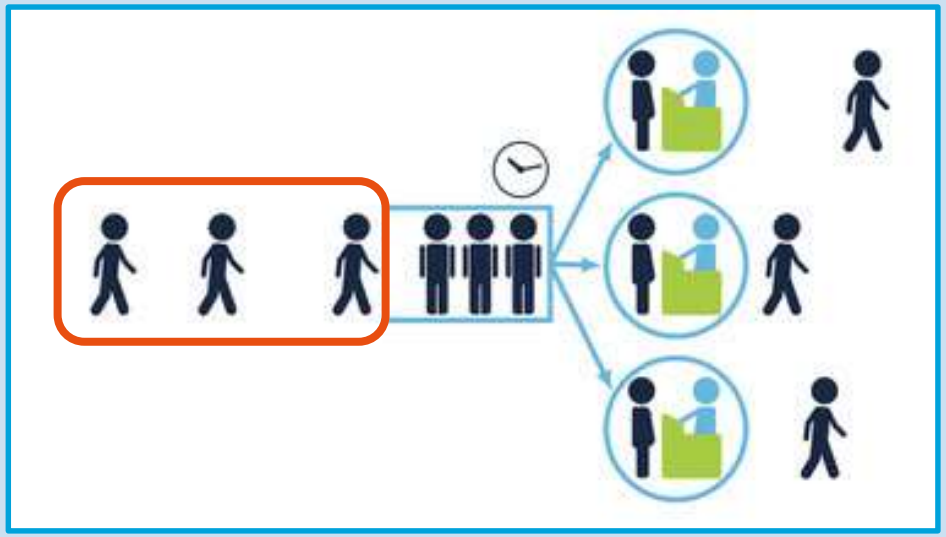
- So, for each day, we are drawing from the Poisson distribution.
- In fact, this fits what is called a Poisson process.

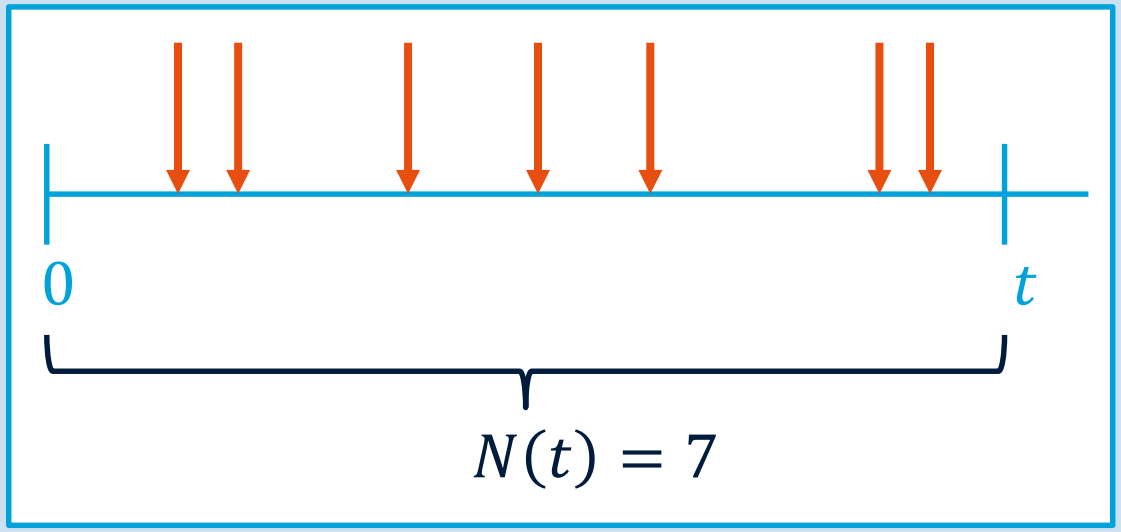


Poisson Process & Distribution

Counting Process

A stochastic process, $\{N(t), t \geq 0\}$, that represents the number of 'events' up to time t is called a ***counting process***.

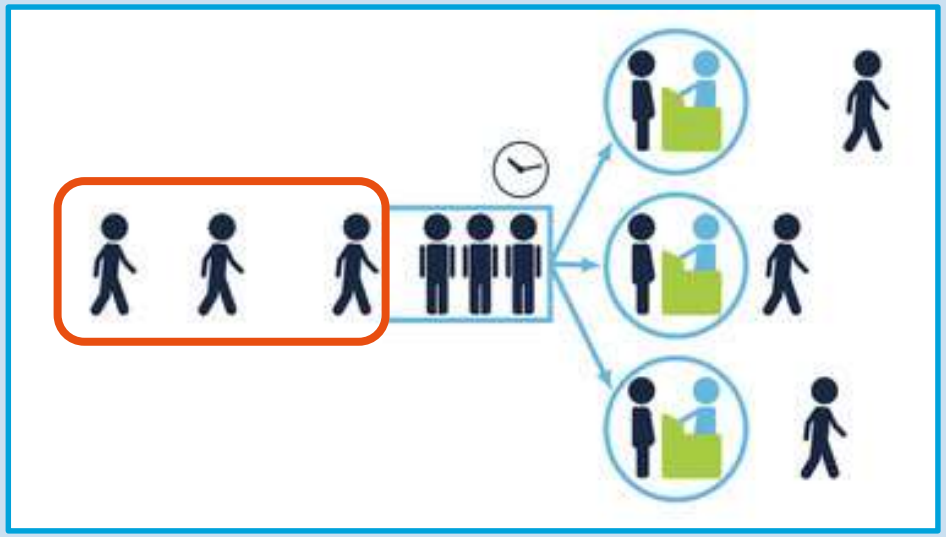


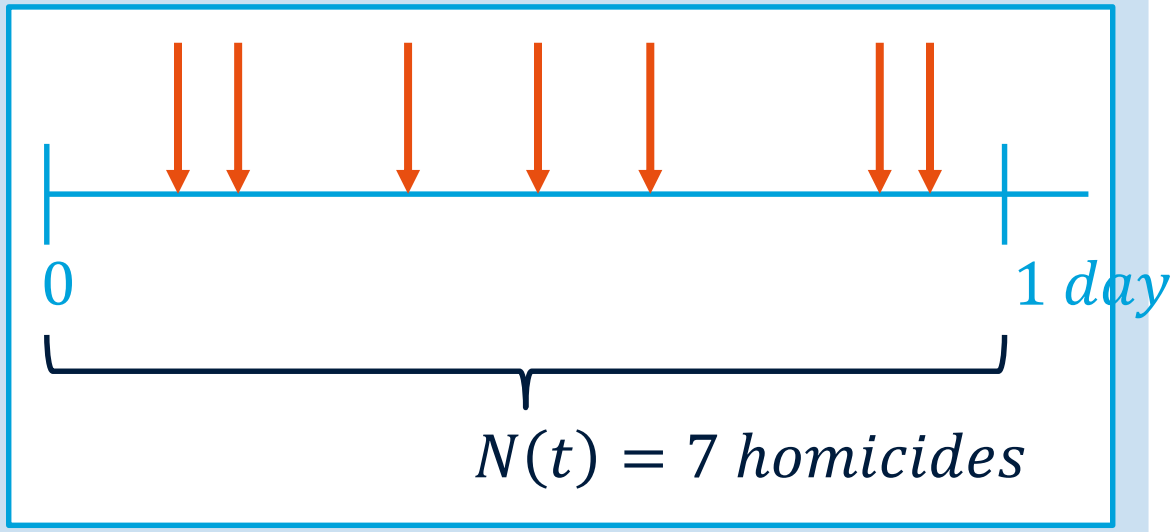


Poisson Process & Distribution

Counting Process

A stochastic process, $\{N(t), t \geq 0\}$, that represents the number of 'events' up to time t is called a ***counting process***.





Poisson Process & Distribution

Counting Process

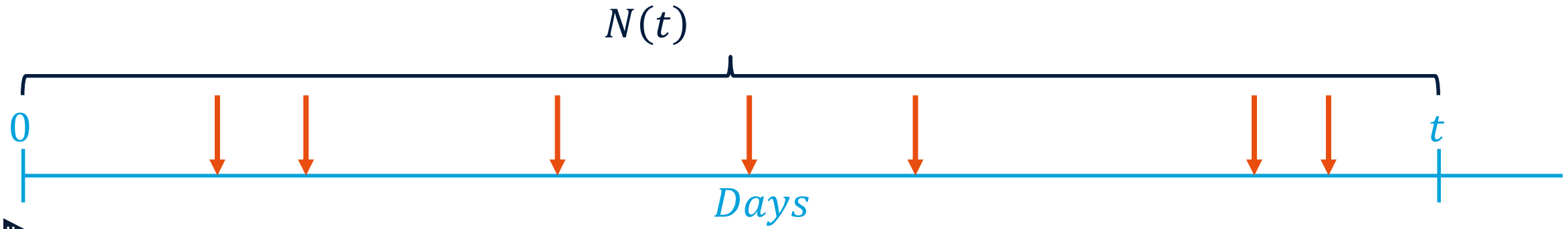
A ***counting process***, $\{N(t), t \geq 0\}$, satisfies:

i. $N(t) \geq 0$. (count up to time t)

ii. $N(t)$ is integer valued.

iii. If $s < t$ then $N(s) \leq N(t)$ (counts are increasing)

iv. For $s < t$, $N(t) - N(s)$ equals the number of 'events' that occur in the interval $(s, t]$

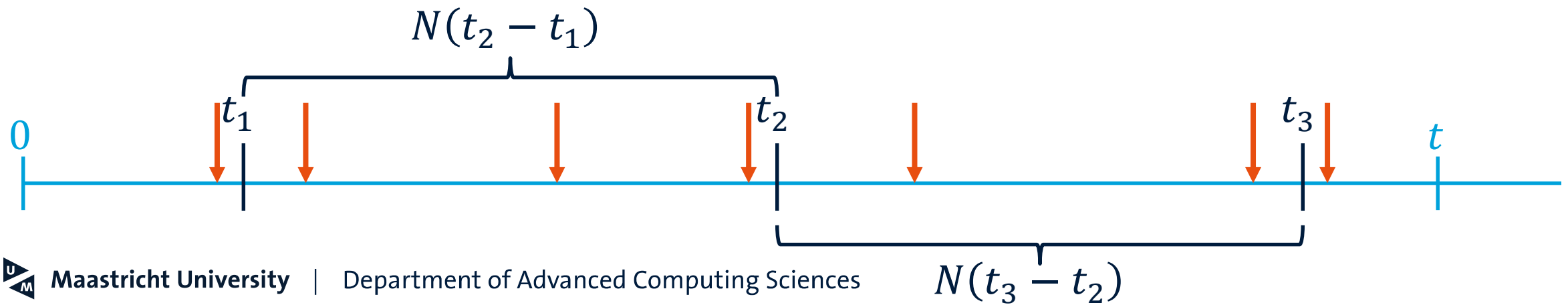


Poisson Process & Distribution

Properties of Counting Process

A ***counting process***, $\{N(t), t \geq 0\}$, has

independent increments if, given $t_1 < t_2 < t_3$, $N(t_2 - t_1)$ and $N(t_3 - t_2)$ are independent.



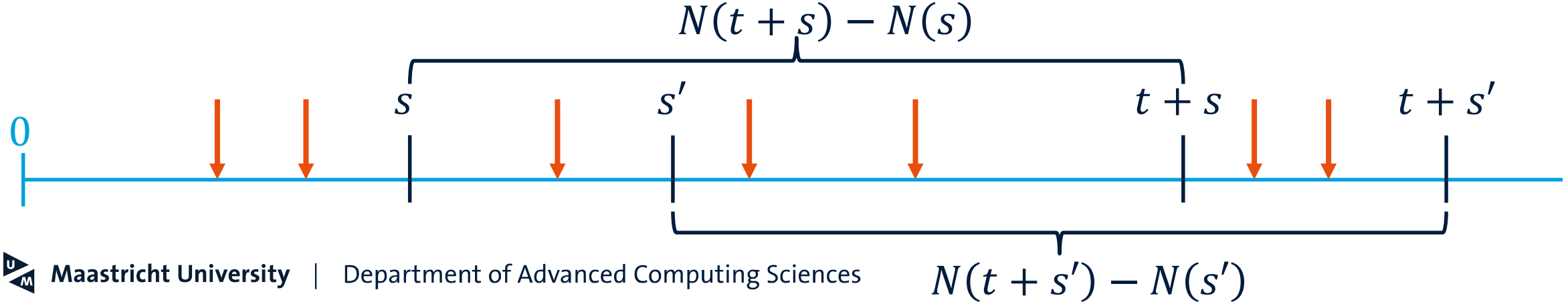
Poisson Process & Distribution

Properties of Counting Process

A ***counting process***, $\{N(t), t \geq 0\}$, has

independent increments if, given $t_1 < t_2 < t_3$, $N(t_2 - t_1)$ and $N(t_3 - t_2)$ are independent.

stationary increments if the distribution of $N(t+s) - N(s)$ is independent of s.



Poisson Process & Distribution

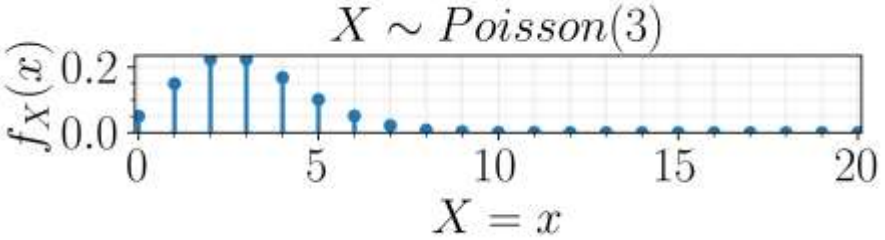
Poisson Process (denoted by $PP(\lambda)$)

A counting process $\{N(t), t \geq 0\}$ is said to be a ***Poisson Process*** with rate $\lambda, \lambda > 0$, if

I. $N(0) = 0$.
II. The process has ***stationary*** and ***independent increments***.
III. Arrivals occur at separate moments in time, i.e. there are ***no simultaneous arrivals*** at any time t.

The number of 'events' during a Poisson Process is captured by the probability distribution called the ***Poisson distribution***.

Poisson Process & Distribution



Poisson Distribution

If we have a ***Poisson Process*** with rate λ and we define X to be the number of arrivals in the interval $[0, t)$, then $X \sim Poi(\lambda t)$ and

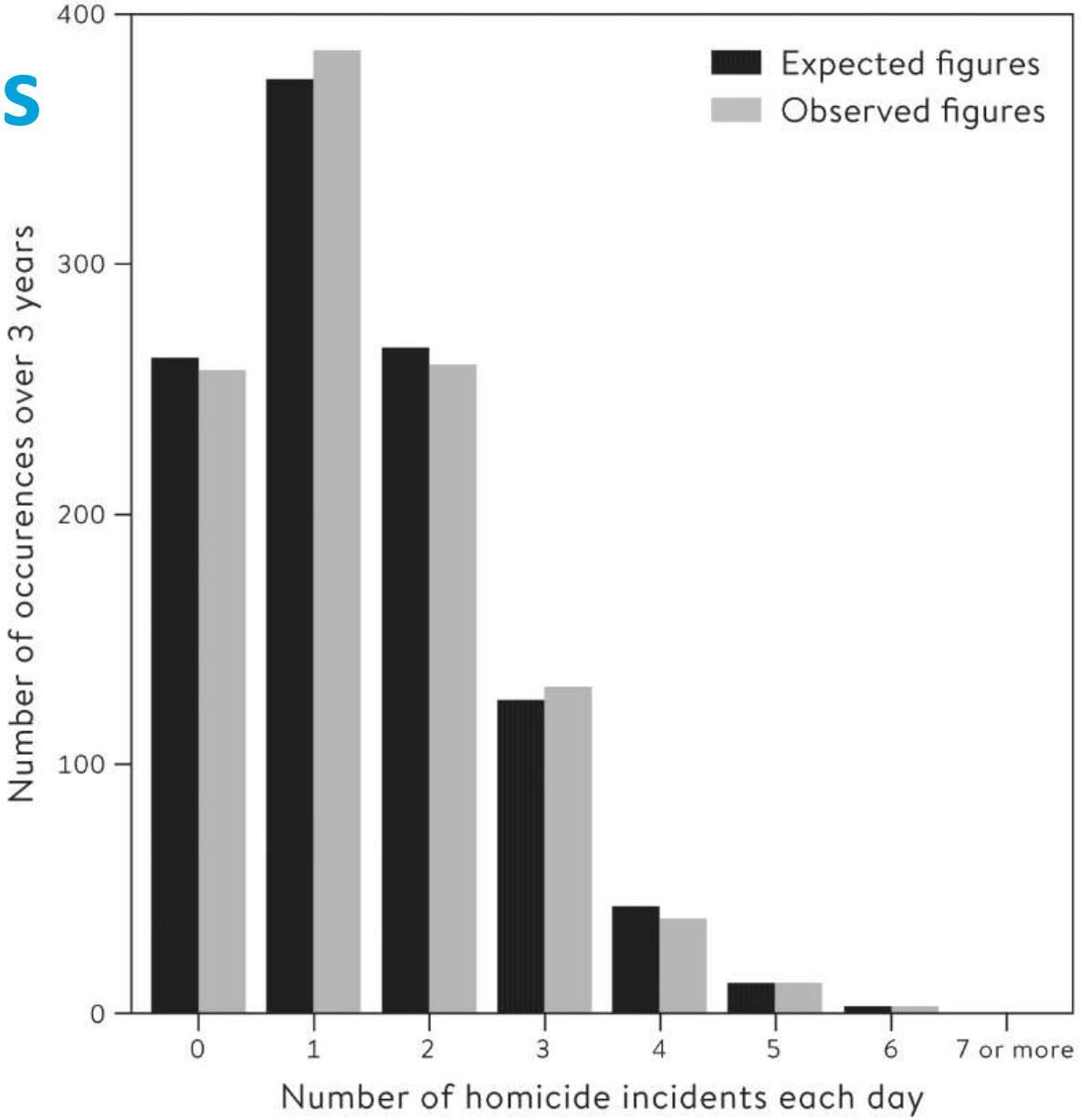
$$P(X = x) = e^{-\lambda t} \frac{(\lambda t)^x}{x!}, \qquad x \in \mathbb{N}$$

Descriptive statistics: $X \sim Poi(\lambda t)$

$$E[X] = \lambda t, Var(X) = \lambda t$$

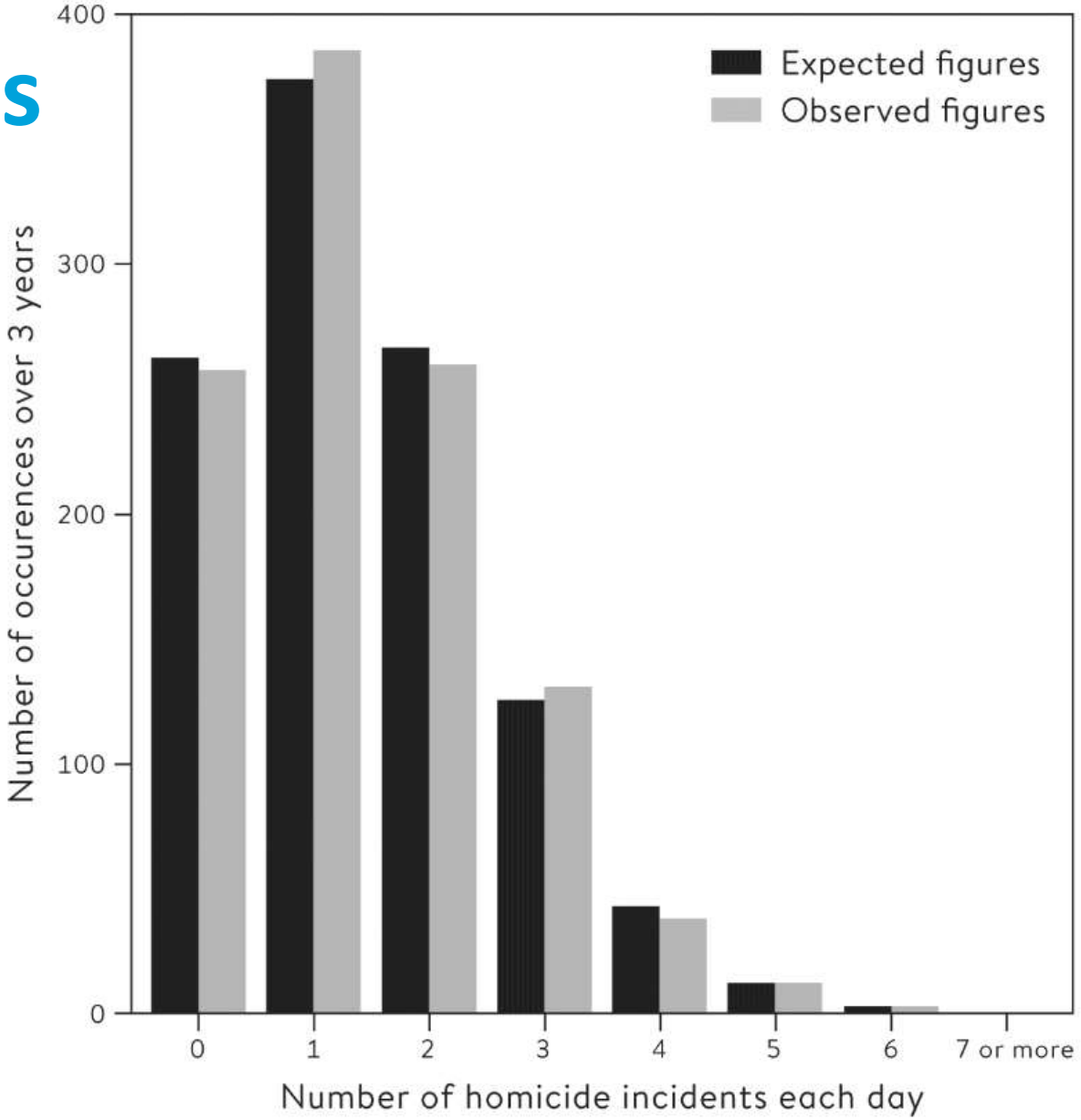
Discrete Random Variables

- This Poisson distribution comes from the assumption that the mean is 1.41 events/day.
- We can then calculate a probability distribution which we multiply by the total count to get the histogram in black.
- In grey is the observation.



Discrete Random Variables

- Based on this, the P(5 homicides) = P(X = 5) = 0.011134.
- Over 1095 days that is an expected frequency of 12 days.
- The actual number of days was 13 so a very good match.
- You will learn to test whether a distribution fits the data well in a future class.



Why do we want a distribution?

- We could have used the natural frequencies directly right?
- There is a branch of statistics that tries to do this: non-parametric statistics.
- But. Having the distribution enables more complicated modeling, and comparisons across countries, for example.

Continuous Random Variables

Example:

$S = [0,1]$, $\mathrm{X}(s) = s$, all outcomes are “equally likely” = **Uniform distribution**.

Cumulative distribution (denoted by $F(x)$)

$$F(x) = P(X \leq x)$$ (is well defined this way)

Continuous Random Variables

Probability density function (denoted by $f(x)$)

$$f(x) = F'(x)$$

Example

$$F(x) = \begin{cases} 0 & if\ x \leq 0, \\ x & if\ 0 \leq x \leq 1, \\ 1 & if\ x \geq 1 \end{cases}$$

$$f(x) = \begin{cases} 0 & if\ x \leq 0, \\ 1 & if\ 0 \leq x \leq 1, \\ 0 & if\ x \geq 1 \end{cases}$$

Continuous Random Variables

Properties of $F(x)$ and $f(x)$

Continuous Random Variables

Example: a) Show that f is indeed a density function

X is an RV with density $f(x) = \begin{cases} \frac{7-2x}{2}, & 2 < x < 3 \\ 0, & \text{elsewhere} \end{cases}$

Continuous Random Variables

Example: b) Find $F(x)$

X is an RV with density $f(x) = \begin{cases} \frac{7-2x}{2}, & 2 < x < 3 \\ 0, & \text{elsewhere} \end{cases}$

=

Continuous Random Variables

Example: c) Find $P\left(1\frac{1}{2} \leq X \leq 2\frac{1}{2}\right)$

X is an RV with density $f(x) = \begin{cases} \frac{7-2x}{2}, & 2 < x < 3 \\ 0, & \text{elsewhere} \end{cases}$

$$F(x) = \begin{cases} 0 & if\ x < 2, \\ -\frac{1}{2}x^2 + \frac{7}{2}x - 5 & if\ 2 \leq x \leq 3, \\ 1 & if\ x > 3 \end{cases}$$

Summaries of Random Variables

- When we take the mean of homicide events, then we are averaging over all the outcomes from random variables for each day.
- We can also calculate other statistics here: median, standard deviation and so on.
- And these statistics are in fact also random variables coming from their own probability distributions (a.k.a the “sampling distribution”).

The mean

- Mean is also called the expectation – though the ‘expectation’ has a clearer mathematical definition in probability.
- Why is it called expectation? It is the most expected amount for a random variable, at least for symmetric random variables.
- How is the mean distributed for the proportion of left handed individuals in the world?

Mathematical Expectation

Expectation (Mean) of a Random Variable

The mean of a Random Variable X is defined as

$$E[X] = \mu_X = \begin{cases} \sum_{x \in S_X} x \cdot f(x) = \sum_{x \in S_X} x \cdot P(X = x) & X \text{ is a discrete RV} \\ \int_{-\infty}^{\infty} x \cdot f(x)\, dx & X \text{ is a continuous RV} \end{cases}$$

Example (Discrete RV)

X: Outcome of a single die roll. $E[X] = \sum_{x \in S_X} x \cdot P(X = x) = \sum_{x \in \{1,2,3,4,5,6\}} \frac{x}{6} = \frac{7}{2}$

Mathematical Expectation

Example (Continuous RV)

X is a continuous RV with PDF

$$f(x) = \begin{cases} \dfrac{7-2x}{2}, & 2 < x < 3 \\ 0, & elsewhere \end{cases}$$

What is $E[X]$? ***Solution:***

$$E[X] = \mu_X = \int_{-\infty}^{\infty} x \cdot f(x)\, dx =$$

Mathematical Expectation

Example (Continuous RV)

X is a continuous RV with PDF

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

What is $E[X]$? ***Solution:***

$$E[X] = \mu_X = \int_{\infty} x \cdot f(x)\, dx$$

Variance of a Random Variable

Variance

Let X be an RV with $\mu_X = E[X]$. Then the variance of X is given by

$$Var(X) = E[(X - E[X])^2] = E[(X - \mu_X)^2]$$

Notation: $V(X)$, $Var(X)$, σ^2, and σ_X^2

Alternative Formula

Let X be an RV with $\mu_X = E[X]$. Then the variance of X is given by

$$Var(X) = E[(X - \mu_X)^2]$$

Observation: always non-negative

Mathematical Expectation

Theorem: $E[aX + b] = a \cdot E[\mathrm{X}] + b$

If X and Y are independent RVs, then

$$E[XY] = E[X] \cdot E[Y]$$

Covariance between Random Variables

Intuition

Dependence of realizations of random variables.

Covariance (denoted by $Cov(X, Y)$)

Let X and Y be RVs with $\mu_X = E[X]$ and $\mu_Y = E[Y]$. Then

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Notation: $Cov(X, Y), \sigma_{XY}$

Remark

Covariance can be positive and negative (recall: Var is always nonnegative)

Covariance between Random Variables

Covariance (denoted by $Cov(X, Y)$)

Let X and Y be RVs with $\mu_X = E[X]$ and $\mu_Y = E[Y]$. Then

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Notation: $Cov(X, Y), \sigma_{XY}$

Alternative Formula

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Covariance between Random Variables

Covariance (denoted by $Cov(X, Y)$)

Let X and Y be RVs with $\mu_X = E[X]$ and $\mu_Y = E[Y]$. Then

$$Cov(X, Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y$$

Example: X, Y independent RVs

$$Cov(X, Y) = E[XY] - \mu_X \mu_Y$$

Example: $Y = X$

$$Cov(X, Y) = E[XY] - \mu_X \mu_Y$$

Correlation between Random Variables

Correlation Coefficient (denoted by ρ)

The correlation coefficient of two RVs X and Y is defined as

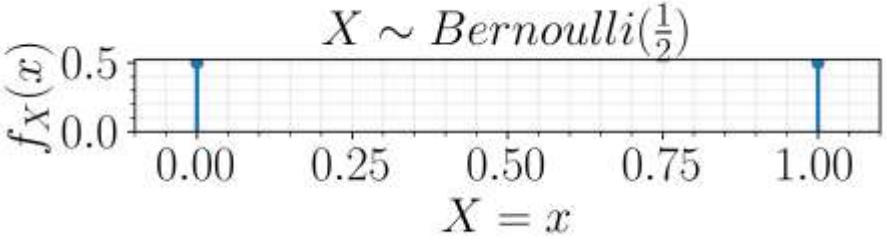
$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E[(X-\mu_X)^2]}\sqrt{E[(Y-\mu_Y)^2]}}$$

Properties of the Correlation Coefficient

Bernoulli Distribution

- Any individual's handedness comes from a discrete probability distribution with some probability for left-handedness and some probability for right-handedness.
- What is the random variable?
- This random variable comes from a Bernoulli distribution, which we have encountered before, for coin tosses. There is a probability *p* on right-handedness and *1-p* on left handedness.

Bernoulli distribution



Bernoulli process

A Bernoulli process is a finite or infinite sequence of independent RV ("trials") $X_1, X_2, X_3, \ldots$, such that each RV (or "trial") follows the distribution

$$p(x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \end{cases}$$

where $x = 1$ denotes "success" and $x = 0$ denotes "failure".

- In a crowd, we expect that if we were shaking hands with one person after another, whether each person was right or left-handed would be a Bernoulli process.

Binomial Distribution

- Suppose now we have a population and we wanted to treat the number of individuals who are left-handed as a random variable.
- Now every time we are measuring what is the number of people who are left-handed, and for each individual there is a population probability of *1-p* of being left handed.
- The random variable “number of left-handers” is described by a Binomial distribution – a generalization of the Bernoulli.

Binomial Distribution

- If we expect x left handers, and we label $q = 1 - p$ as probability of right-handedness, then, if we assume we have first x left handers and then $n - x$ right handers:

  $p^x q^{n-x}$ is the probability

  But we need to account for all the other potential ways of ordering.

Binomial Distribution

$p^x q^{n-x}$ is the probability

But we need to account for all the other potential ways of ordering.

Pretend n = 10 and x = 2. Then we need all the ways that the 2 left-handers could be organized, and all the ways that the 8 right-handers could be organized.

Binomial Distribution

Pretend n = 10 and x = 2. Then we need all the ways that the 2 left-handers could be organized, and all the ways that the 8 right-handers could be organized.
Ways 10 people can be ordered = 10! = 10 * 9 * 8…
Ways 2 people can be ordered = 2! = 2*1
Ways 8 people can be ordered = 8!

We also need to consider that the 2 ways the left-handers organize interacts with the 8! ways that the right handers organize.

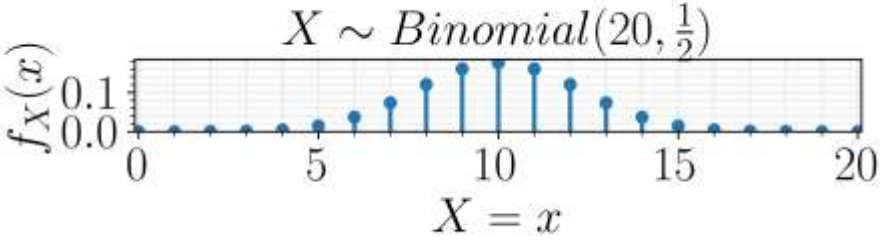
Binomial Distribution

- So, the probability of 2 left handers in a population of 10 is:

$$\binom{10}{2} p^2 q^8$$

$$= \frac{10!}{2!8!} p^2 q^8$$

Binomial distribution

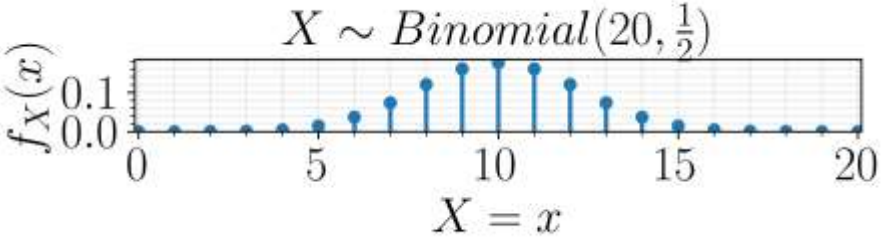


Example

25% of the families in a village have a subscription to a local newspaper. If we ask 10 families if they have a subscription, what the probability that exactly 4 have one?

Bernoulli process

Binomial distribution



Binomial distribution

If we do n independent trials, where each time the success probability is $p \in [0,1]$, and we define X to be the number of successes, then $X \sim B(n,p)$,

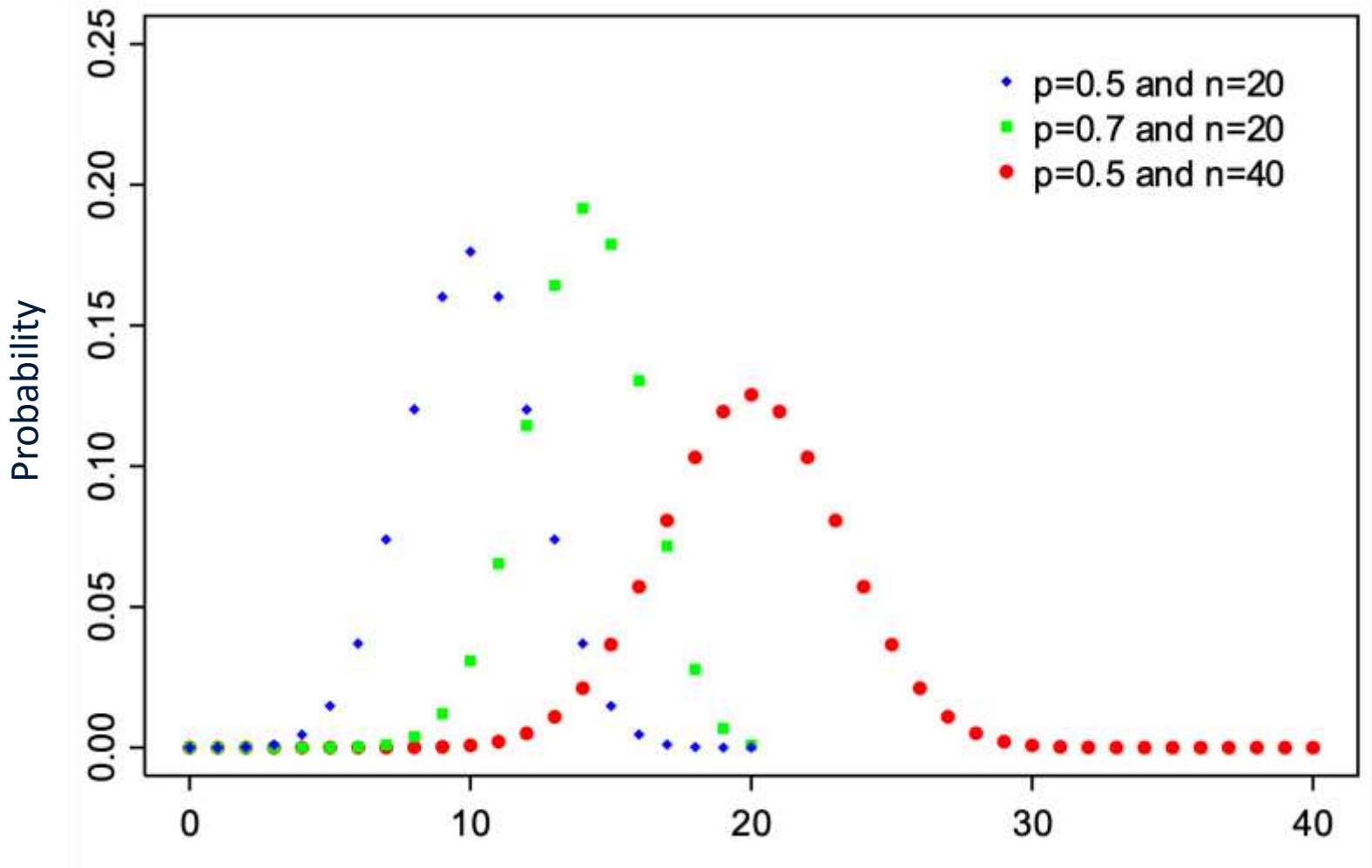
pick x out of n | x successes each w.p. p | $n-x$ failures each w.p. $1-p$

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},\ x = 0,1,2,\dots,n$$

Descriptive statistics: $X \sim B(n,p)$

$$E[X] = np$$
$$Var(X) = np(1-p)$$

Binomial Distribution



Mean of Binomial Distribution

- Now if we wanted a margin of error on the mean, we need to know its *sampling* distribution. We could do this using *bootstrapping: sampling with replacement and recalculating the statistic of interest*.

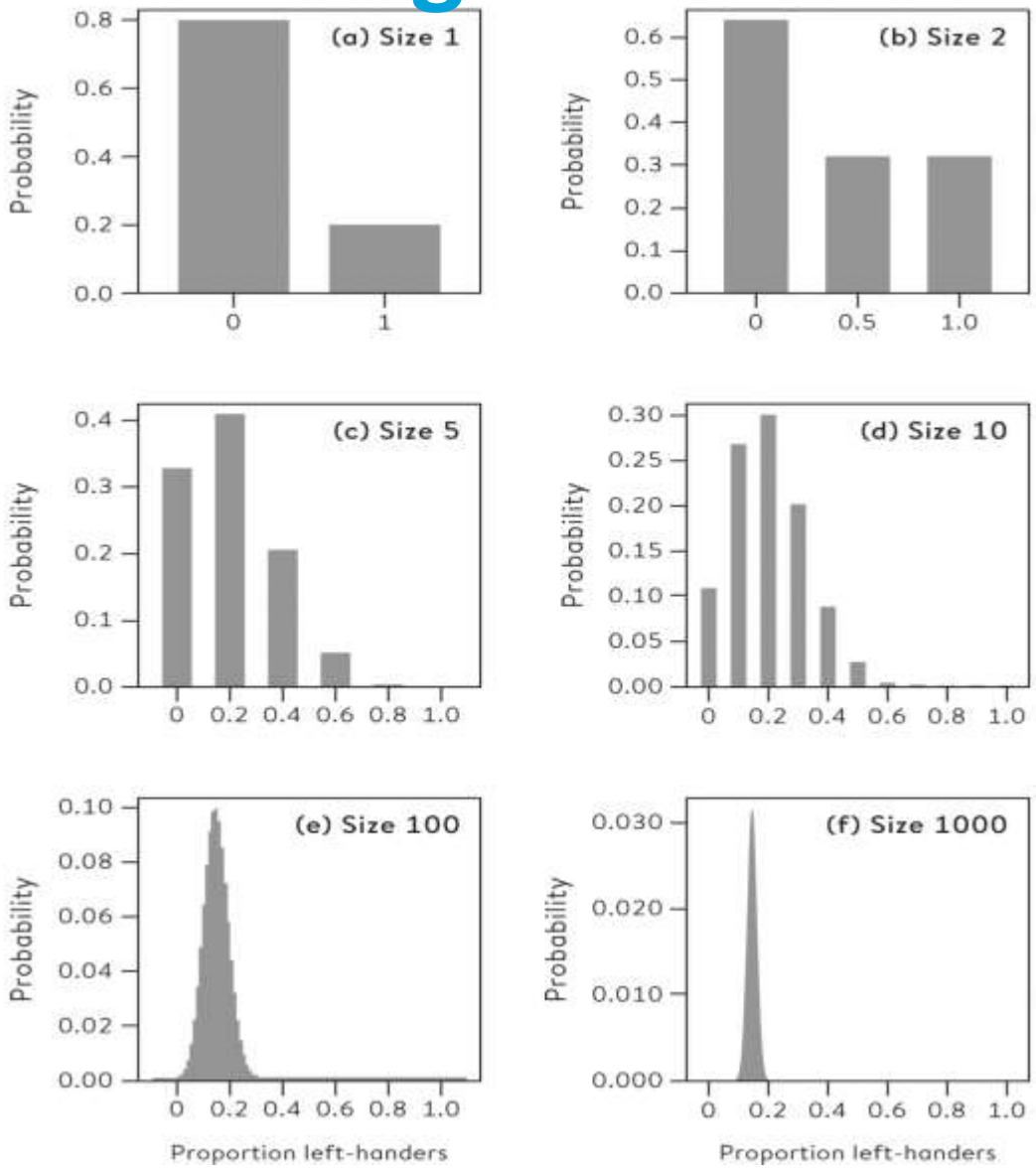
- But we can actually also do it using the sampling distribution.

Distributions of Summary Statistics

- Every summary statistic when derived from the realizations of random variables has an associated distribution.
- But why bother trying to figure this out when we can bootstrap all of it?
- Bootstrap is harder with large datasets, larger models and sometimes you have variables where using a distribution is the better choice.

Mean of Binomial Distribution: Margin of Error

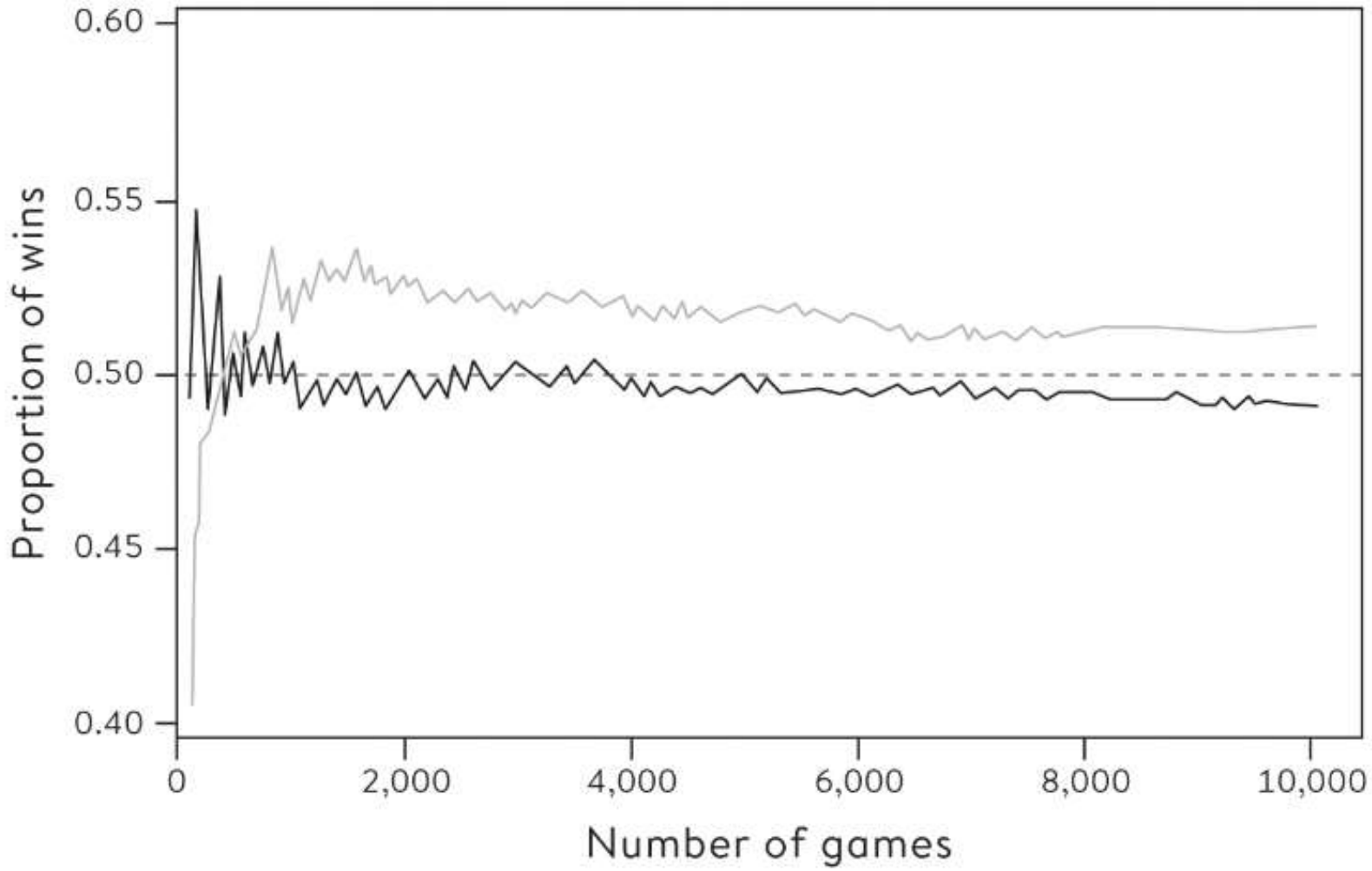
- We can see how the distribution looks for different numbers of people in our survey.
- As the number increases, the shape of the distribution seems to look like something we've seen before…



Law of large numbers

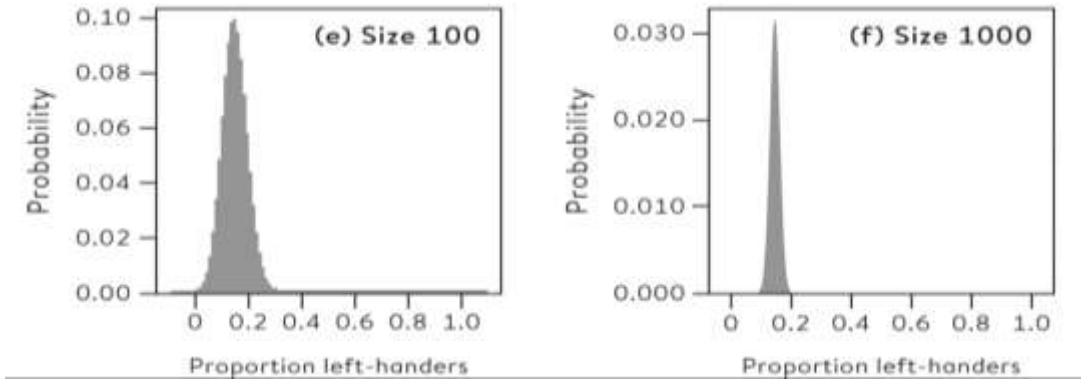
- First step: Law of large numbers
- The law of large numbers states that as your sample size *n* gets larger and larger, then the sample mean (the estimate) gets closer and closer to the population mean (the parameter).
- You will have to wait for a proof of this.
- But you've already seen it in action

Law of large numbers



Central Limit Theorem

- The law of large numbers says that the sample mean converges to the population mean as the *n* increases.
- We don’t know anything about the distribution of the sample mean… but it looks familiar.



Central Limit Theorem

- Indeed, the central limit theorem proves that:

- For distributions without very heavy tails, the sample mean converges to the normal [bell curve] distribution with the population mean as its mean (law of large numbers) and standard deviation = $\sigma/\sqrt{n}$ = population standard error!
- The implications of this are huge!

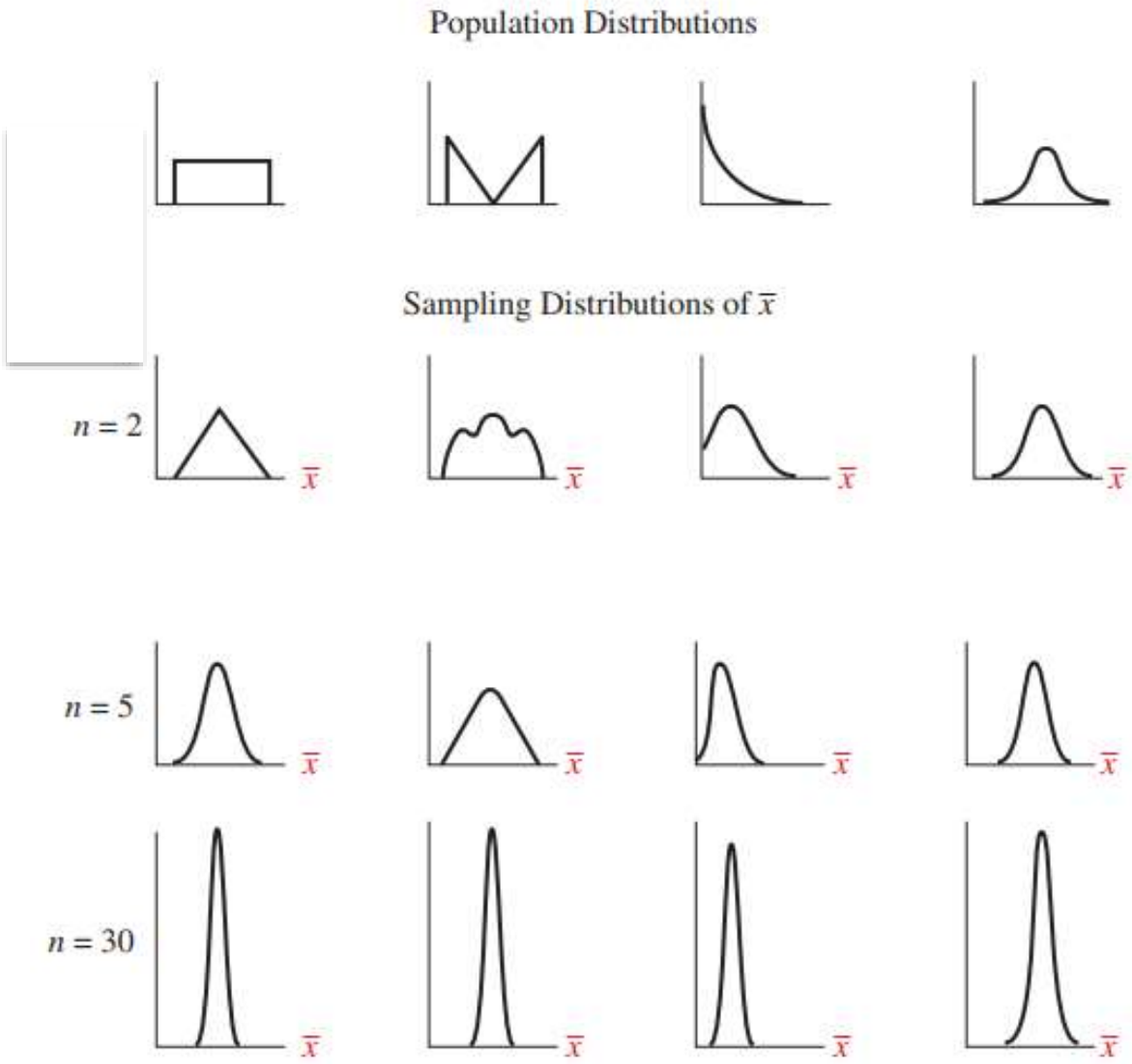
Central Limit Theorem

- We can always pretend remember than some datapoint is a random variable comes from a distribution.
- The central limit theorem says that the mean of that random variable, WHATEVER it is, converges to the true population mean and a normal distribution.
- The fact that this is true is incredible and spurred some poeticism from Francis Galton.

Central Limit Theorem

The law [central limit theorem] would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway.

-- Francis Galton



Assumptions of CLT

1. All random variables are independent.
2. All random variables come from the same distribution.
3. The population variance of the random variable is finite and non-zero.

Why the normal distribution?

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

- Turns out that when you have sums of random variables they lead to a symmetric distribution.
- Sums of normally distribution R.V.s are normally distributed.
- Since sums of random variables leads to a symmetric distribution and gaussian sums lead to gaussians, sums of *any* random variables give a normal.

Why the normal distribution?

- See the following videos for more visual introductions to how the central limit theorem works and why a normal distribution:
  - https://www.youtube.com/watch?v=zeJD6dqJ5lo&t=0s
  - https://www.youtube.com/watch?v=d_qvLDhkg00

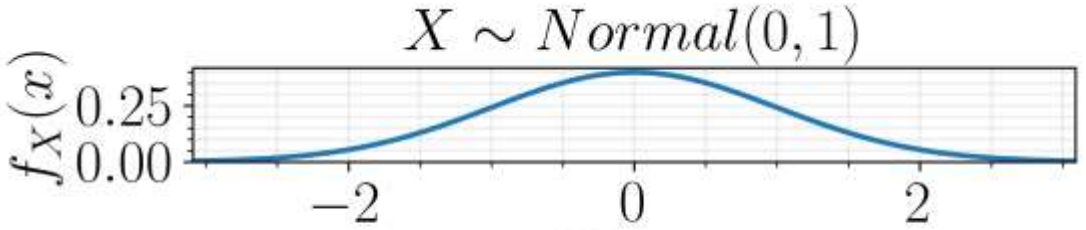
Central Limit Theorem

- SO. What can we do with the central limit theorem?

- Define the probability for the mean and through that a sense for uncertainty/margins or error!

- Now that we know that the mean converges to the normal distribution, we can estimate the 95% range within which we expect the population parameter to fall!

Normal distribution



Normal Distribution

An RV X has a normal distribution with parameters μ and σ, $X \sim N(\mu, \sigma)$ if the density is given by

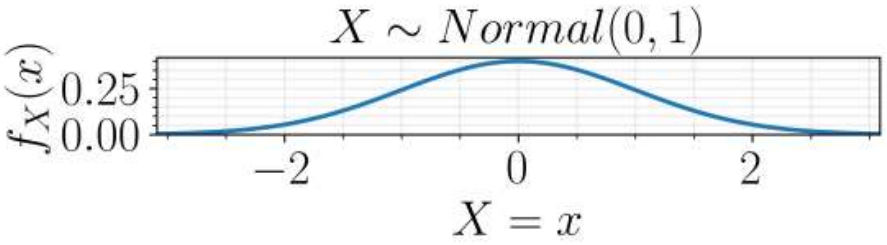
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \text{if } x \in \mathbb{R}$$

Descriptive statistics: $X \sim N(\mu, \sigma)$

f is symmetric around and maximal for $X = \mu$

$$E[X] = \mu$$
$$Var(X) = \sigma^2$$

Normal distribution



Theorem: Standard Normal Distribution

If $X \sim N(\mu, \sigma)$ and $Z = \frac{X-\mu}{\sigma}$, then $Z \sim N(0,1)$ (called standard normal dist.)

Proof:

$$z = \frac{x-\mu}{\sigma}, dz = \frac{1}{\sigma}dx$$

Confidence Interval

- If we can calculate the sampling distribution for the mean, then we can assess the probability of different values for the population parameter.

- How best to characterize this distribution?

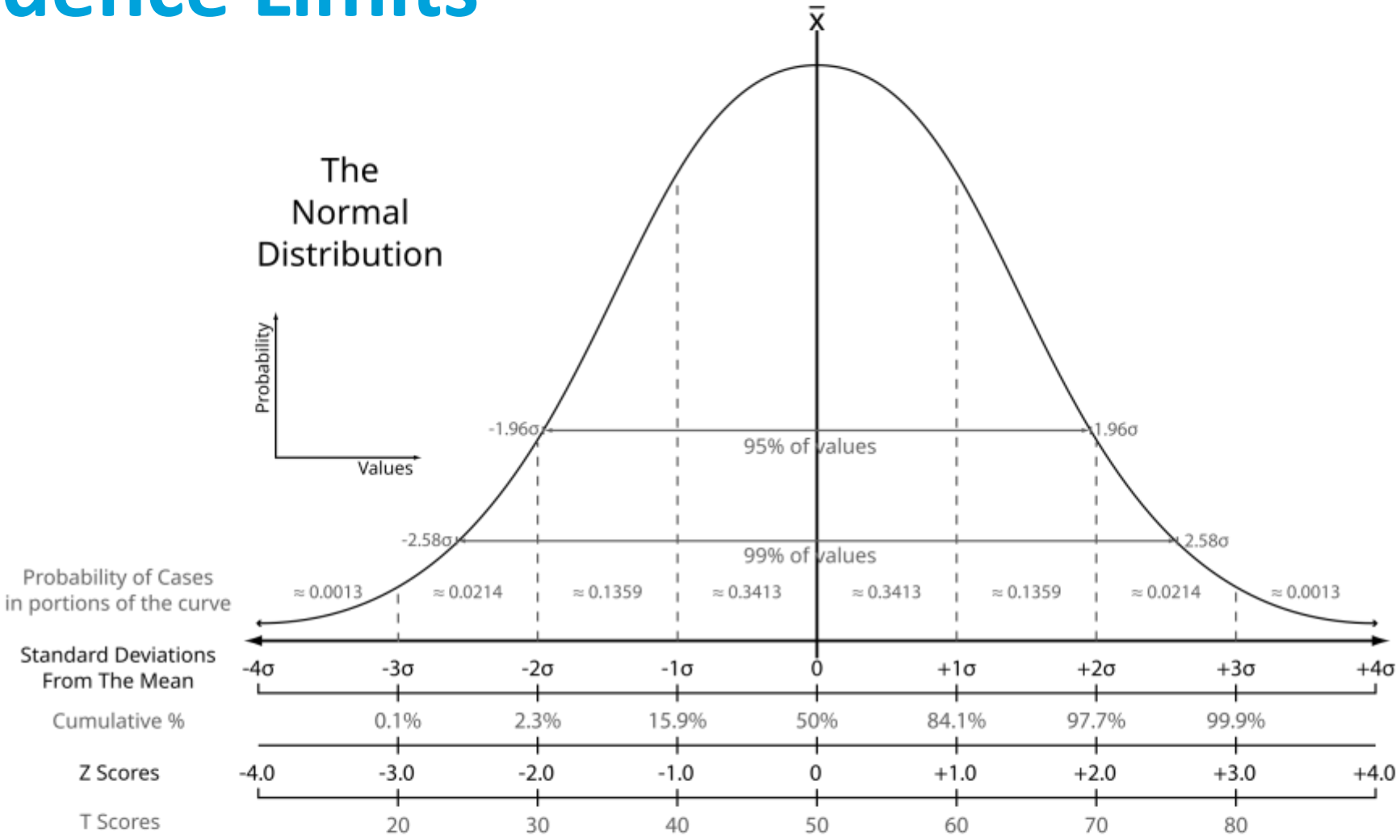
Confidence Interval

- Brings us back to quantifying uncertainty.
- Since sampling distribution for a mean under CLT is a symmetric distribution, the standard deviation is a great measure to use.
- Also, we know based on the standard normal what amount of standard deviations gives us 95% of possible values.

Standard Normal Distribution

- Every normal distribution can be transformed into the standard normal.
- How? Simple variable substitution.
- $Z = \frac{X - \mu}{\sigma}$, $Z \sim N(0, 1)$
- This means for any random variable distributed as $N(\mu, \sigma)$, if we wanted the probability for a range $[x_1, x_2]$, we can use P($[\frac{x_1 - \mu}{\sigma}, \frac{x_2 - \mu}{\sigma}]$) under a standard normal.

Confidence Limits



Sampling Distribution for a Mean

- Assuming central limit theorem holds:

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right), \boldsymbol{\frac{\sigma}{\sqrt{n}}\ called\ Standard\ Error}$$

- If we want the range within which, 95 times out of a 100, our sample mean will cover the population mean, we calculate:

$$\bar{x} \pm 1.96 * STD_{\bar{x}} = [\bar{x} - 1.96 * \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 * \frac{\sigma}{\sqrt{n}}]$$

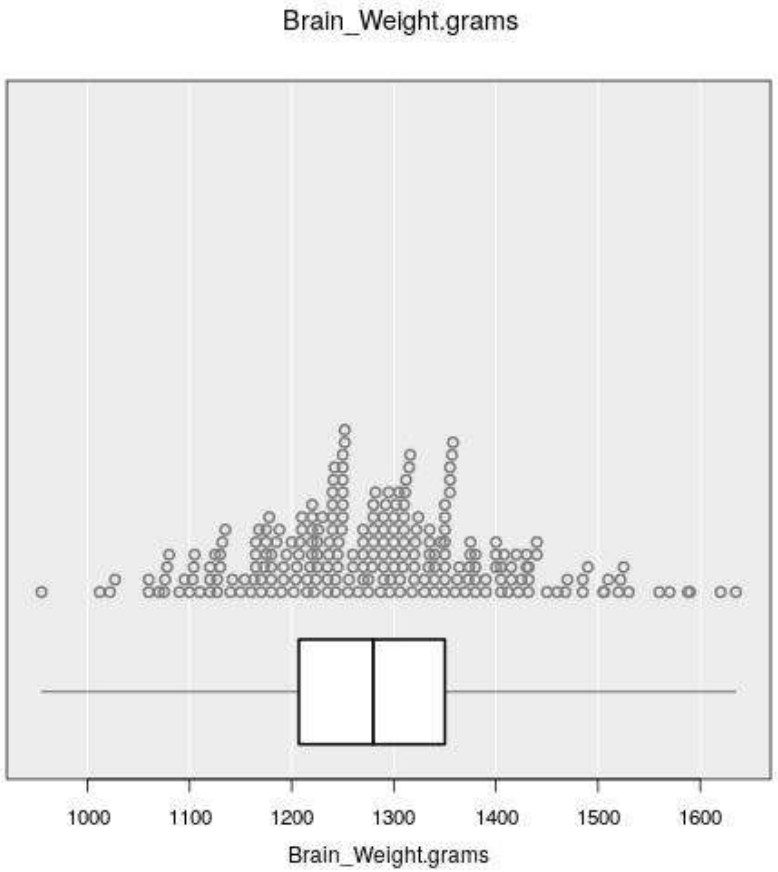
Central Limit Theorem

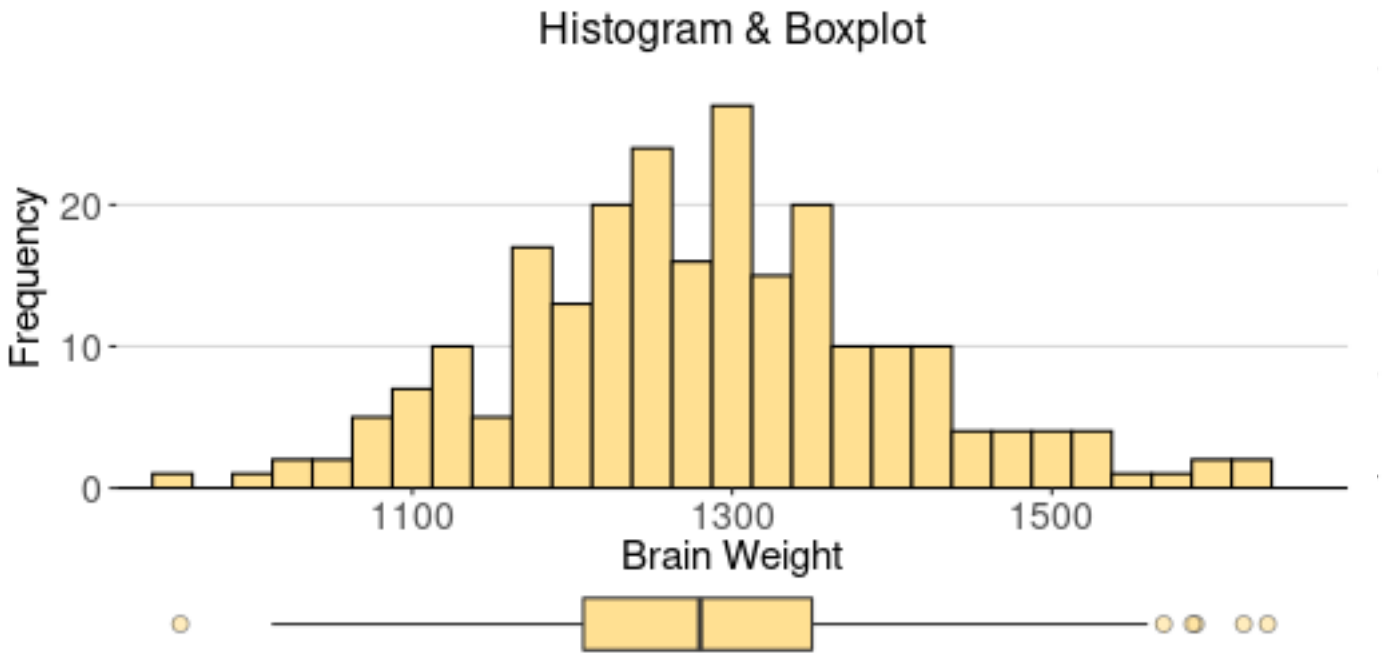
- States that a sample mean for random variables from almost any distribution converges to the normal distribution.
- That is, regardless of distribution of random variable:

$$\lim_{n \to \infty} \bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Confidence Intervals for a Single Mean

- Say we had a set of brain weights for babies.





Confidence Intervals for a Single Mean

- We can estimate the confidence intervals for the sample mean by calculating the standard error.
- Standard error = $\hat{\sigma}/\sqrt{n}$
- 95% confidence intervals are: $\bar{x} \pm 1.96 * \hat{\sigma}/\sqrt{n}$
- $z = \frac{x-\mu}{\sigma}$ is called a z-score. For example, z = 1.96 for a probability of 0.95.

Confidence Intervals for a Single Mean

95% confidence intervals are: $\bar{x} \pm 1.96 * \hat{\sigma}/\sqrt{n}$

Where does the 1.96 come from?

The Central Limit Theorem allowing us to assume a normal distribution for the sample mean!
Under the normal, 1.96 standard errors = 95% probability. 1.96 is referred to as $z_{\frac{\alpha}{2}}$ for $\alpha = 0.05$.

Confidence Intervals for a Single Mean

Descriptive Statistics:

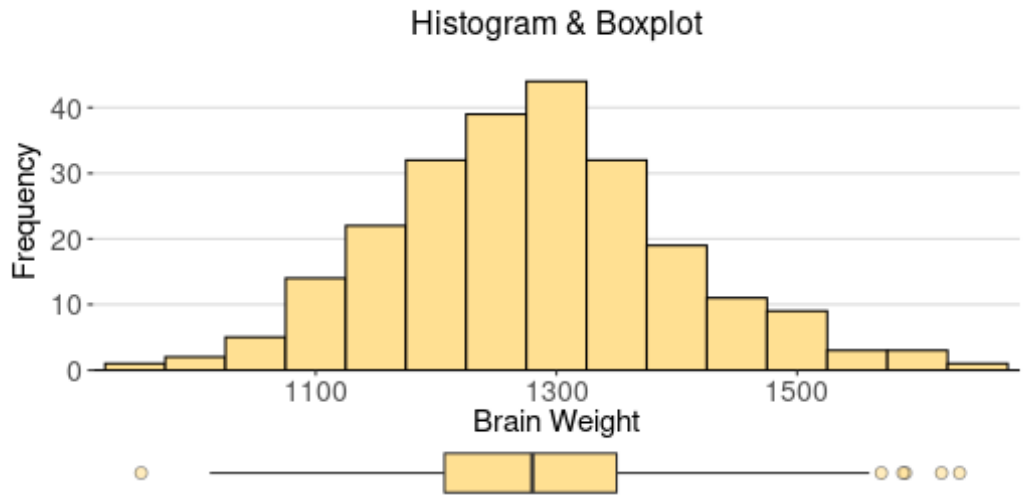
Sample Size	Sample Mean	Sample Standard Deviation
237	1,283	120.3

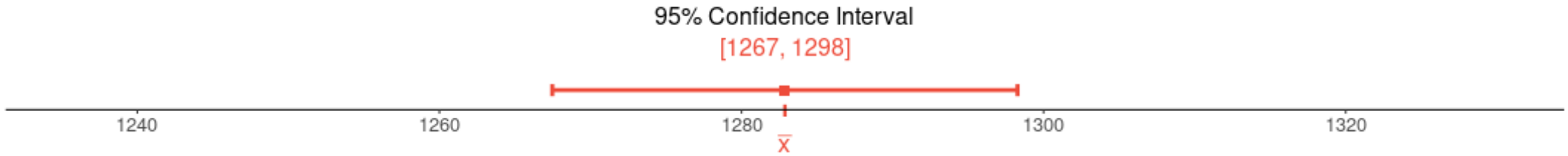
Estimate of Population Mean:

Point Estimate	Standard Error	Margin of Error
1,283	7.817	± 15.4

Confidence Interval:

Population Parameter	Lower Bound	Upper Bound	Confidence Level
Mean μ	1,267	1,298	95%





Sampling Distribution for a Mean

Sampling Distribution for a Mean

Sampling Distribution for a Mean

Sampling Distribution for a Mean

Confidence Intervals for Small Datasets

- Say we want to get the margin of error for the sample mean.
- We will still rely on the standard error:

$$se = \frac{std}{\sqrt{n}}$$

- $Mean\ \bar{x}\ \pm (something) * se$
- Before the something was $z_{\frac{\alpha}{2}}$ = 1.96 ($\alpha = 0.05$)

Confidence Intervals for Small Datasets

- Number of hours watching television in USA

N	Mean	StDev	SE Mean	95% CI
1324	2.9800	2.6600	0.0731	(2.8366, 3.1234)

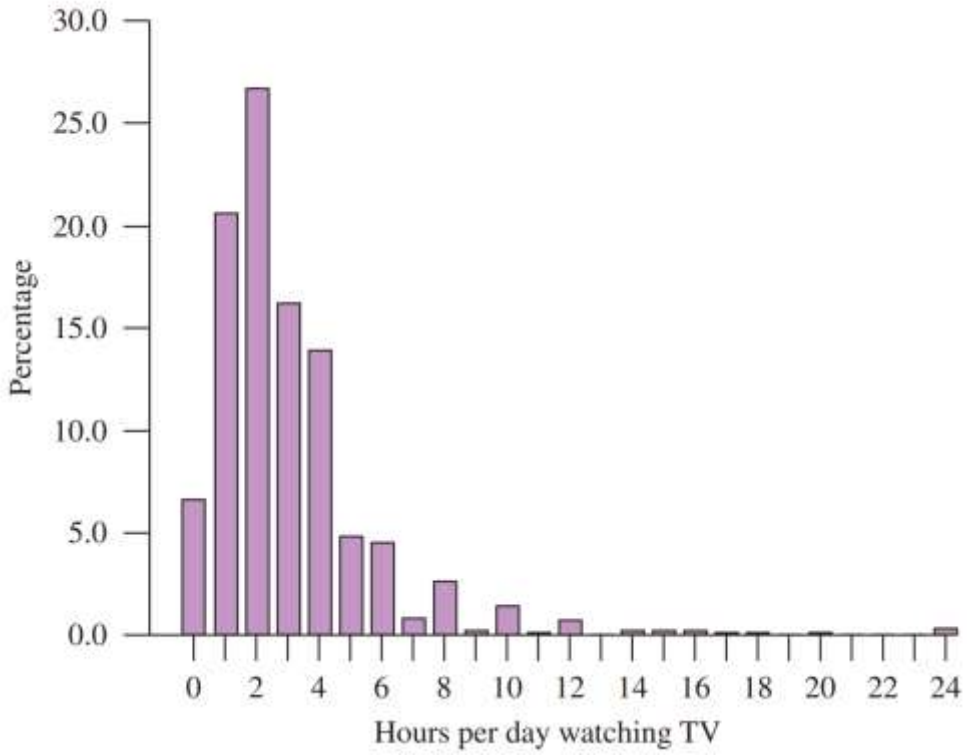
Standard Error = $\frac{s}{\sqrt{n}} = \frac{2.66}{\sqrt{1324}} =$

0.0731

95% confidence interval =

[2.84, 3.12] hours

Large sample size – more confidence in mean.



Confidence Intervals for Small Datasets

- For big datasets, we can assume the mean is close to a normal distribution and use standard error.
- For small datasets, this is unlikely.
- Let's say we can at least assume the X is normally distributed.

Confidence Intervals for Small Datasets

- When trying to estimate the standard error for a small dataset, we need to estimate the standard deviation, which can have more error (farther from population parameter) as a result.
- To make up for this, rather than use a z-score $(z_{\frac{\alpha}{2}})$ we instead use a t-score that comes from a bell-like distribution with thicker tails.

t-Distribution

- When trying to estimate the standard error for a small dataset, we need to estimate the standard deviation, which can have more error (farther from population parameter) as a result.
- To make up for this, rather than use a z-score $(z_{\frac{\alpha}{2}})$ we instead use a t-score that comes from a bell-like distribution with thicker tails.

t-Distribution Definition

t-Distribution

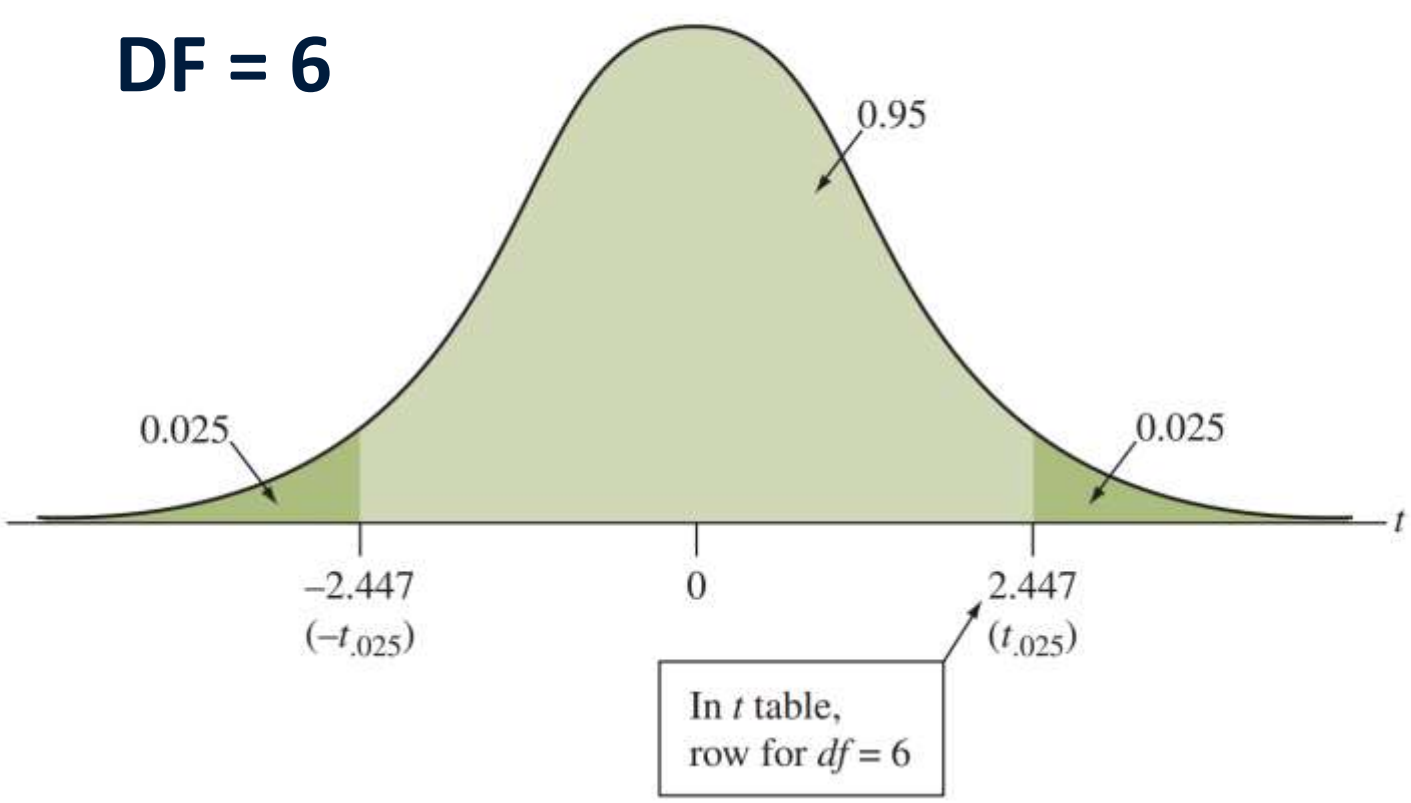
- t- Distribution is bell-shaped and for a mean of 0, is symmetric about 0.
- Probability depends on the number of samples which guides the degrees of freedom (df = samples -1) of the distribution.
- Has thicker tails and more variability as a result than the normal.
- Can get a better sense for confidence in smaller samples, that is, we can be more sure our sample parameter estimate and interval captures the population parameter estimate using the $t - score$ multiplied by the standard error.

t-Distribution Definition

t-Distrib

The Probab
incredibly c
how the dis
numbers of

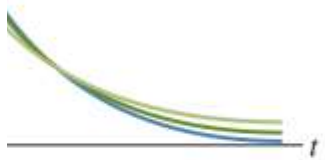
Once the n
the t-distrib
the normal



ard normal distribution
/ariability than *t* distribution)

tribution, *df* = 6

istribution, *df* = 2

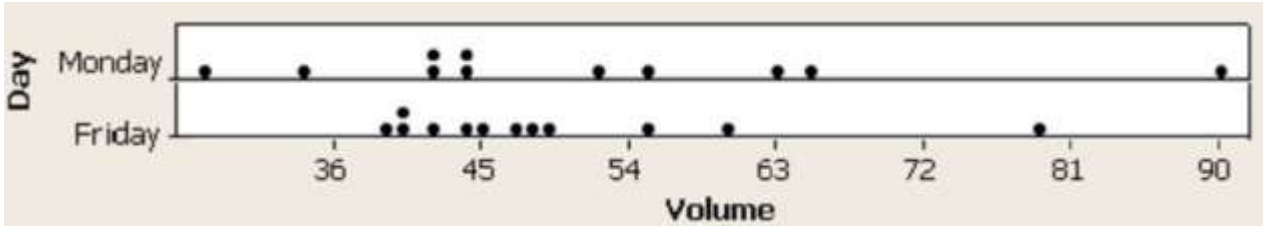


for *t* and for
rd normal

t-Distribution

- Small scale stock market activity.

Day	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Monday	11	51.82	17.19	29.00	43.00	45.00	64.00	91.00



- What is the confidence interval for Monday?
- $t_{\frac{0.05}{2}}^{10} = 2.228$

Is it fair to assume a Normal distribution?

Is it fair to assume a normal distribution?

- Using the t-distribution is actually robust.

Robust Statistical Method

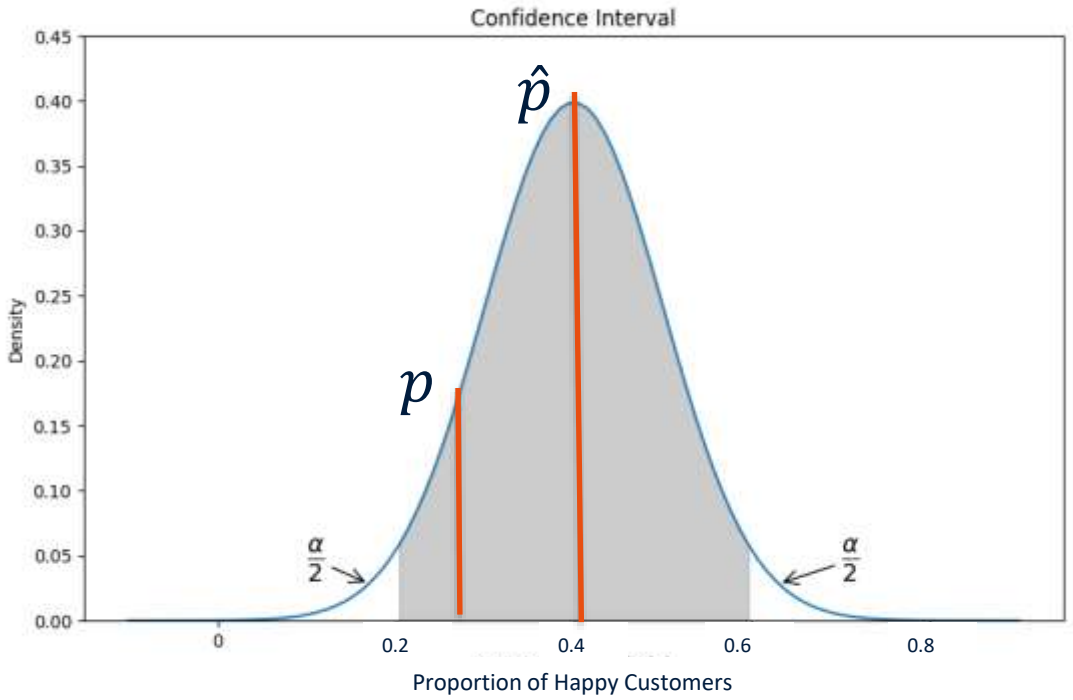
A statistical method is said to be robust with respect to a particular assumption if it performs adequately even when that assumption is modestly violated.

- (1) you can't tell if you have a normal distribution when you have small samples and
- (2) even if it isn't, the t-Dist. gives confidence intervals that tends to capture the population.

When doesn’t the t-distribution work?

- Data aren’t independently generated.
- Data has some extreme outliers.

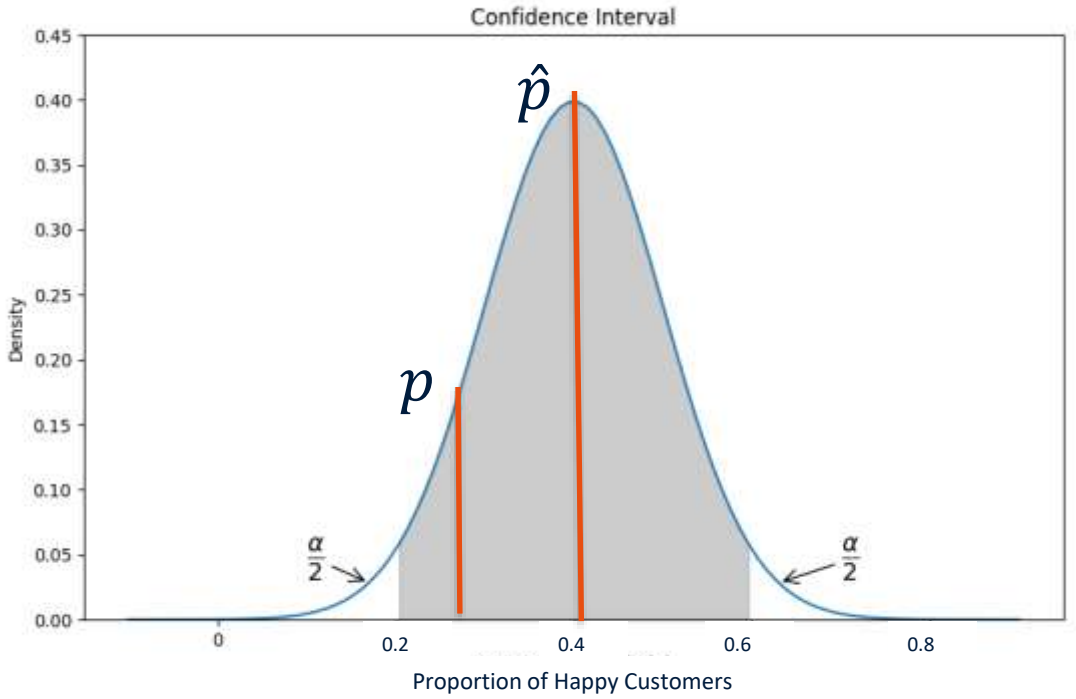
Confidence Interval for Proportion of a Single Sample



- If n is large and $\hat{p}$ is not too close to 0 or 1 ($n\hat{p}, n(1 -$

$$P\left(p \in [\hat{p} \pm z\alpha_{/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]\right) \approx 1 - \alpha$$

Confidence Interval for Proportion of a Single Sample



$$P\left(p \in [\hat{p} \pm z\alpha_{/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]\right) \approx 1-\alpha$$

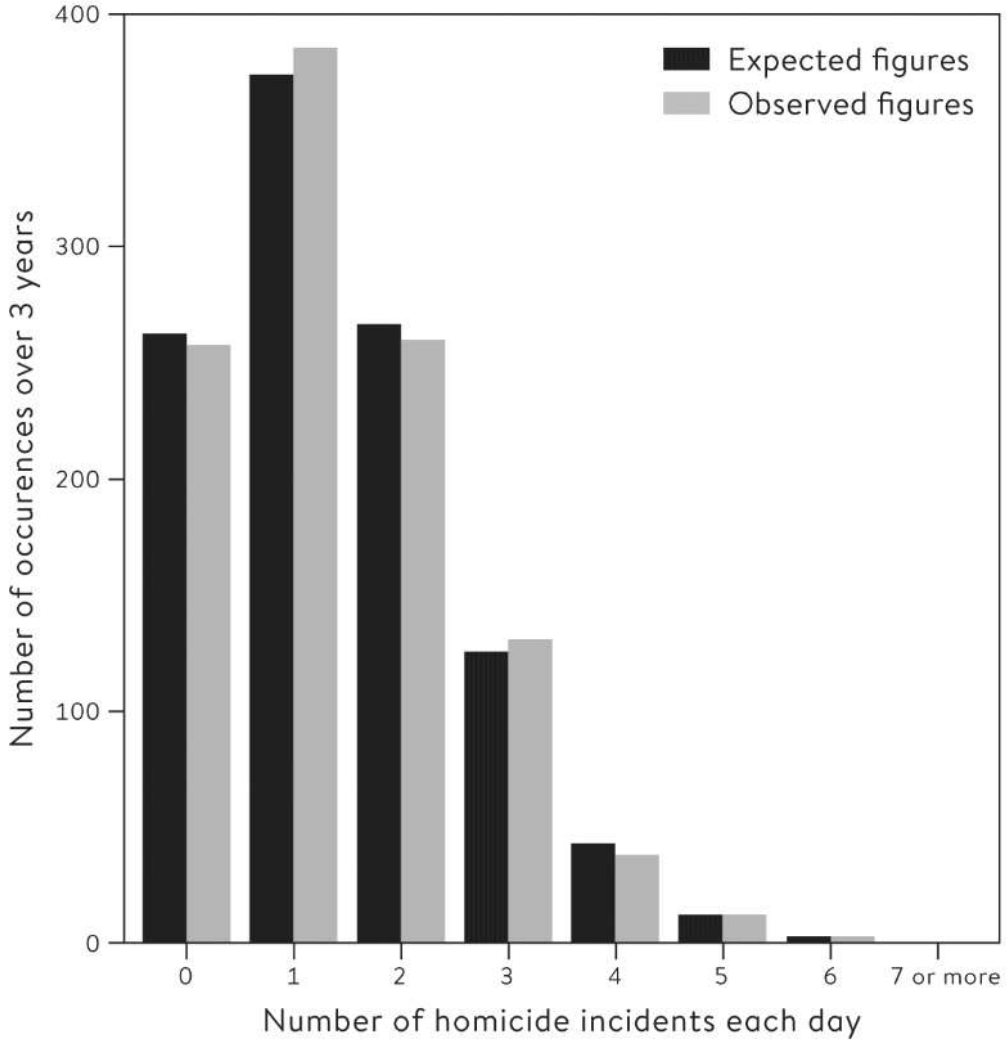
- If in an exit poll we see that 53.8 % of 3889 people voted for one candidate then we get:
- $SD = \frac{\sqrt{\hat{p}(1-\hat{p})}}{n} = \frac{\sqrt{0.538}*\sqrt{0.462}}{3889} = 0.008 \sim 0.01$
- Suggests that CI is 51.6 to 55.8% for $\alpha = 0.05$.

Sample Size Needed

- The general formula of $\hat{p} \pm 1.96 * se$ applied when you have large random samples.
- For proportions this means you have at least 15 successes AND 15 failures.
- Most cases, this is easily satisfied, when it isn't you can always use bootstrapping.

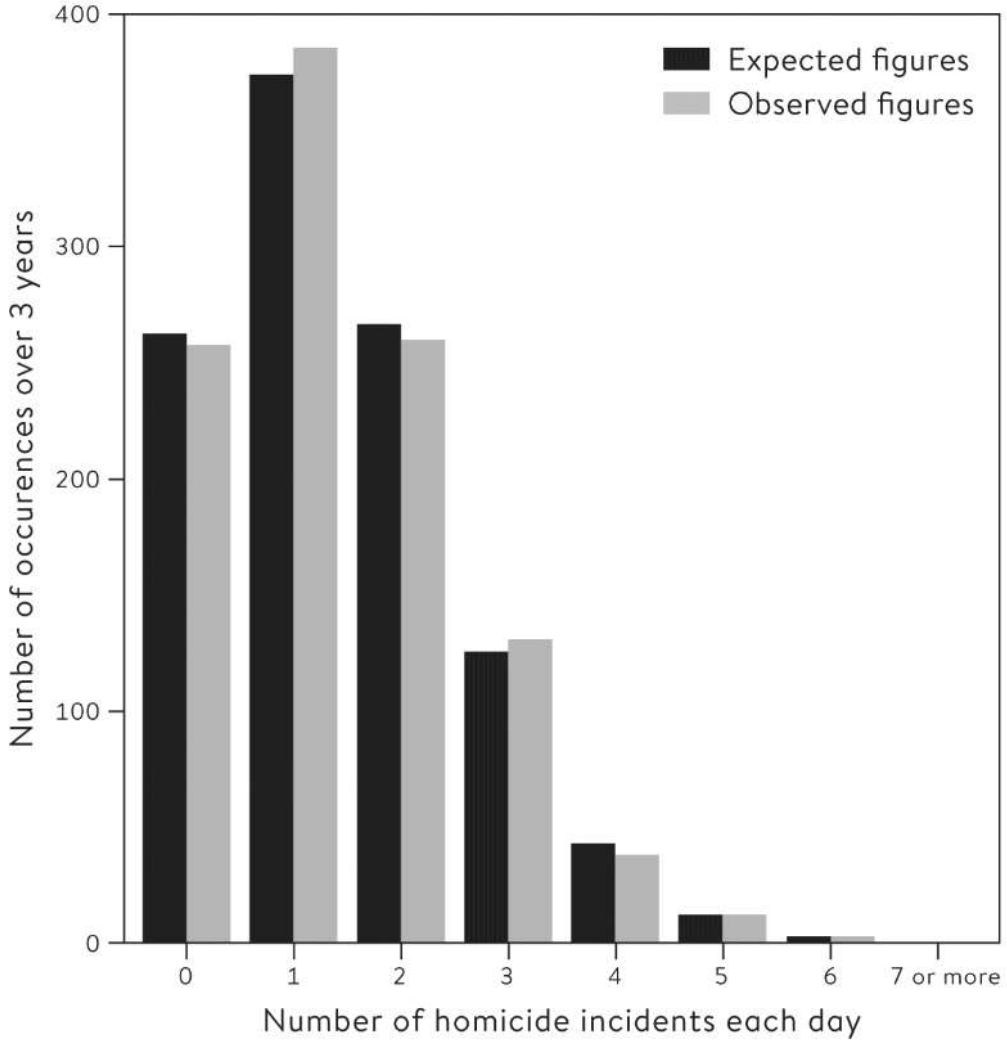
Confidence Interval for a Poisson

- $Mean \pm 1.96 * se$
- Now if we knew that there were 497 homicides in one year (note that we are changing the bin size from days), and 557 the following year, can we judge if 557 is outside 95% CI?

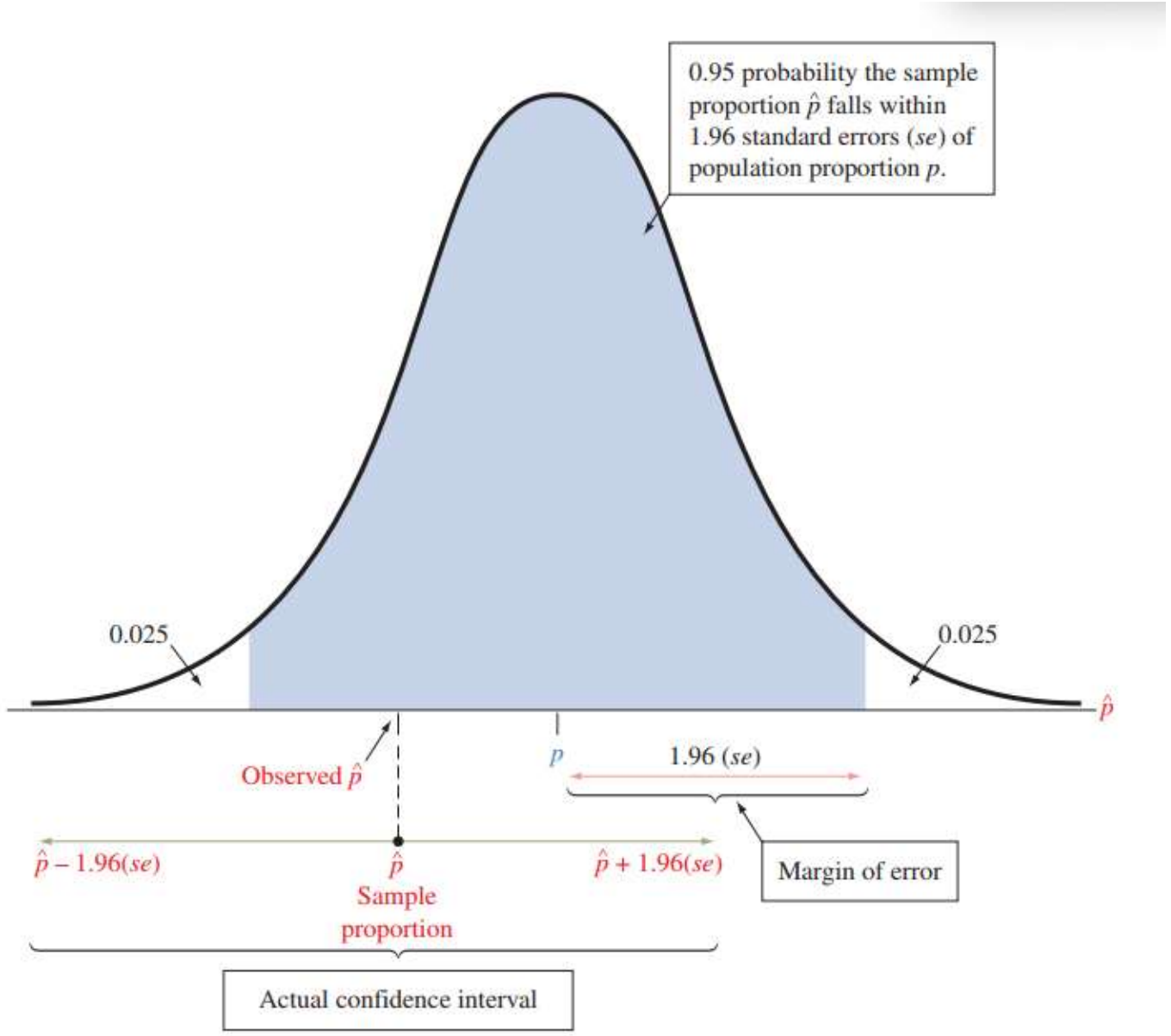


Confidence Interval for a Poisson

- For Poisson, we know the mean and variance are the same.
- So if we say, we are estimating the rate from one year = 497/year.
- Then the variance = 497.
- Standard deviation = Standard error (why?) = $\sqrt{497}$ = 43.7
- So is 557 outside 95% CI?



Confidence intervals



How much should we trust confidence intervals?

- Confidence intervals for the speed of light in the early 19th century did NOT cover the population parameter we now know.
- So why was that? Well confidence intervals cover the following case of quantifying measurement uncertainty:
- Type A uncertainty: Uncertainty that reduces with more samples. But does nothing for:
- Type B uncertainty: Systematic errors that have to do with the nature of measurement itself.