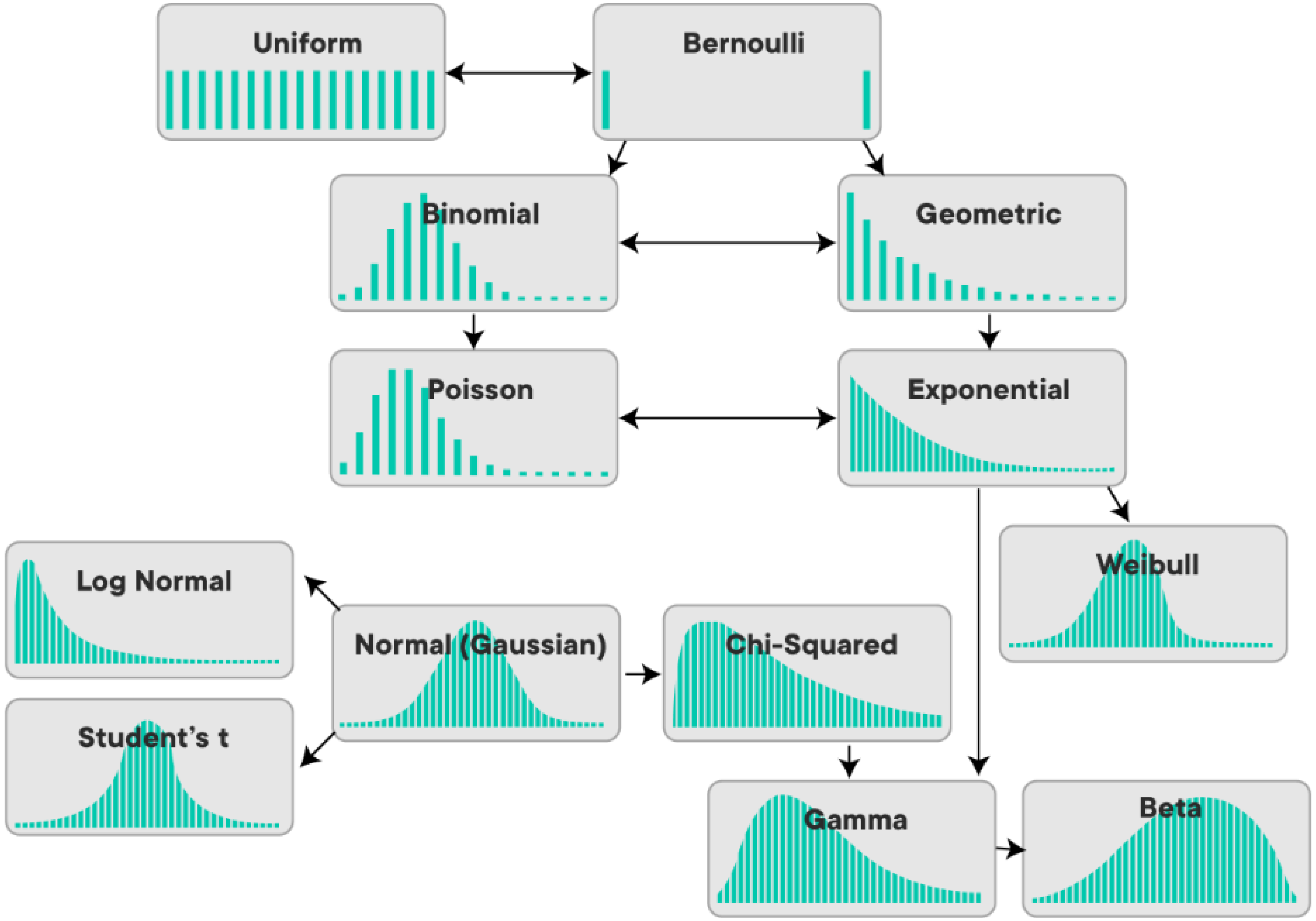
CHAPTER 2

Data and Sampling Distributions

Sections 7 to 12

Data Distribution

- A data distribution is a function or a listing which shows all the possible values (or intervals) of the data. It also (and this is important) tells you how often each value occurs. Often, the data in a distribution will be ordered from smallest to largest, and graphs and charts allow you to easily see both the values and the frequency with which they appear.



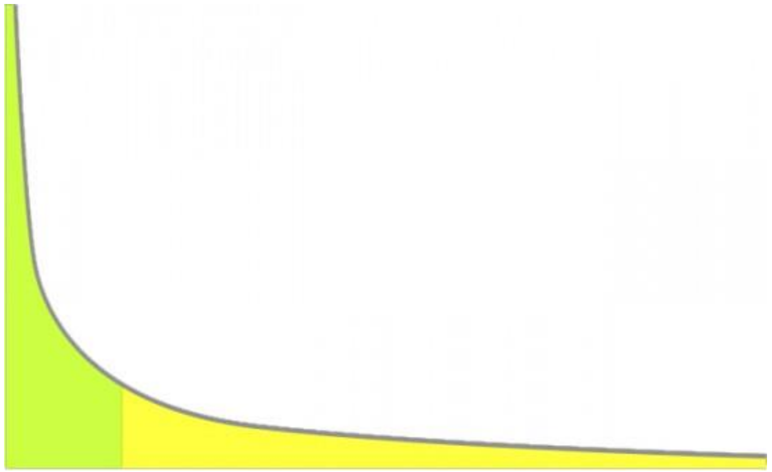
Attention! Normal distribution is not that “NORMAL”

- It is a common misconception that the normal distribution is called that because most data follows a normal distribution—that is, it is the normal thing.
- Most of the variables used in a typical data science project—in fact, most raw data as a whole—are *not* normally distributed.
- The utility of the normal distribution derives from the fact that many statistics *are* normally distributed in their sampling distribution.
- Even so, assumptions of normality are generally a last resort, used when empirical probability distributions, or bootstrap distributions, are not available.

Statistic: Information based on a study of the number of times something happens or is present

Long-tailed distributions

- Data is generally not normally distributed. Sometimes, the distribution is highly *skewed* (asymmetric)
- Long tails are widely recognized in practical work. The fat tails mean that extreme events occur more frequently in reality than would be predicted by the normal distribution.
- Black Swan Theory: the extreme impact of **rare and unpredictable outlier** events—and the human tendency to find simplistic explanations for these events
- Assuming a normal distribution can lead to underestimation of extreme events (“black swans”).



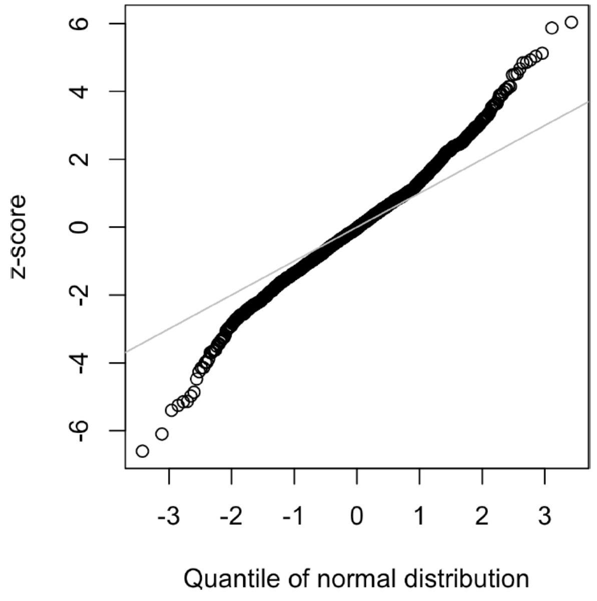
Tail
The long narrow portion of a frequency distribution, where relatively extreme values occur at low frequency.
Skew
Where one tail of a distribution is longer than the other.

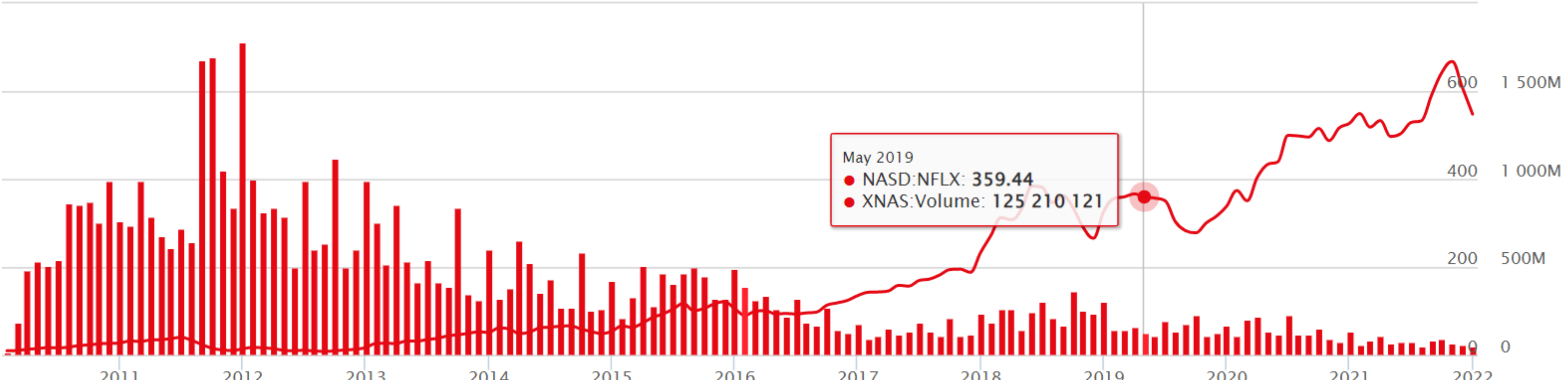
Long-tailed distributions

The Black Swan: The Impact of the Highly Improbable is a 2007 book by author and former options trader Nassim Nicholas Taleb. The book focuses on the extreme impact of rare and unpredictable outlier events—and the human tendency to find simplistic explanations for these events, retrospectively. Taleb calls this the Black Swan theory.

Long-tailed distributions

- A good example to illustrate the long-tailed nature of data is stock returns.





Key Ideas

- Most data is not normally distributed.
- Assuming a normal distribution can lead to underestimation of extreme events (“black swans”).

What is a statistical test?

Statistical tests are used in hypothesis testing. They can be used to:

- Determine whether a predictor variable has a statistically significant relationship with an outcome variable.
- Estimate the difference between two or more groups.

Statistical tests assume a null hypothesis of no relationship or no difference between groups. Then they determine whether the observed data fall outside of the range of values predicted by the null hypothesis.

- Expectation is defined loosely as "nothing unusual or of note in the data" (e.g., no correlation between variables or predictable patterns). This is also termed the "null hypothesis" or "null model"

H0: The sample data follow the hypothesized distribution.
H1: The sample data do not follow the hypothesized distribution.

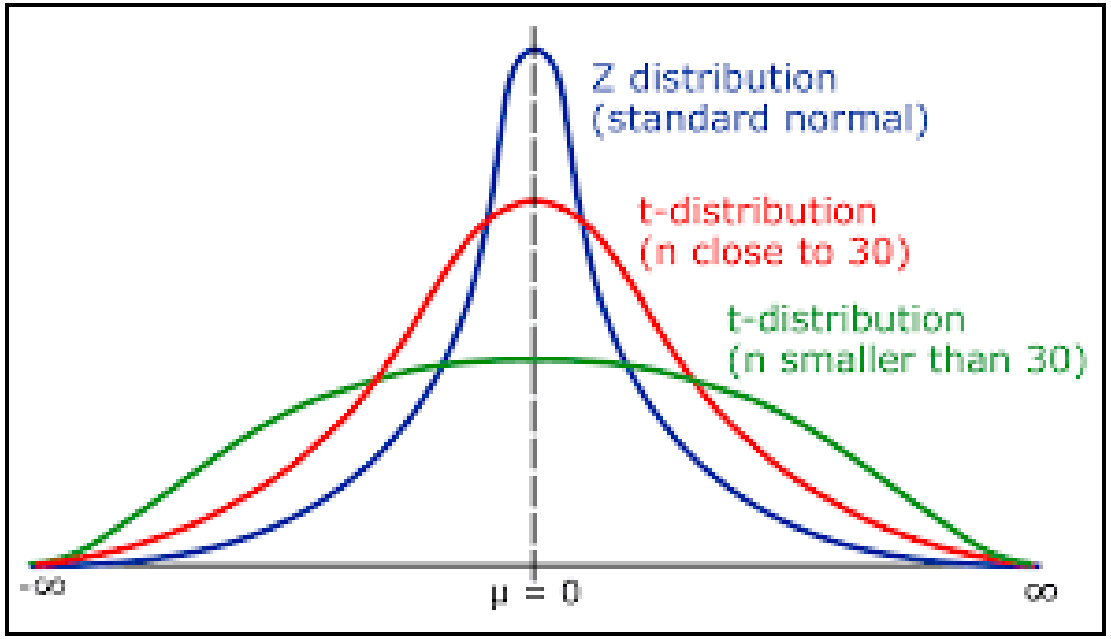
- Computing a statistic gives a p-value, The p-value will tell you if your test results are significant or not

- Compare p-value to alpha: The alpha level(α). This is chosen by you, or the researcher. The usual alpha level is 0.05 (5%), but you could also have other levels like 0.01 or 0.10.

If p-value < alpha, we accept the null hypothesis

Student’s t-Distribution

- The *t-distribution* is a normally shaped distribution, except that it is a bit thicker and longer on the tails.
- Distributions of sample means are typically shaped like a t-distribution
- There is a family of t-distributions that differ depending on how large the sample is.
- The larger the sample, the more normally shaped the t-distribution becomes.
- Degrees of freedom: A parameter that allows the t-distribution to adjust to different sample sizes, statistics, and numbers of groups.
- If the sample size is ***n***, then the t-distribution has ***n-1 degrees of freedom***



$$PDF = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\,\Gamma\left(\frac{n-1}{2}\right)}\left(1 + \frac{x^2}{n-1}\right)^{-\frac{n}{2}}$$

Some history

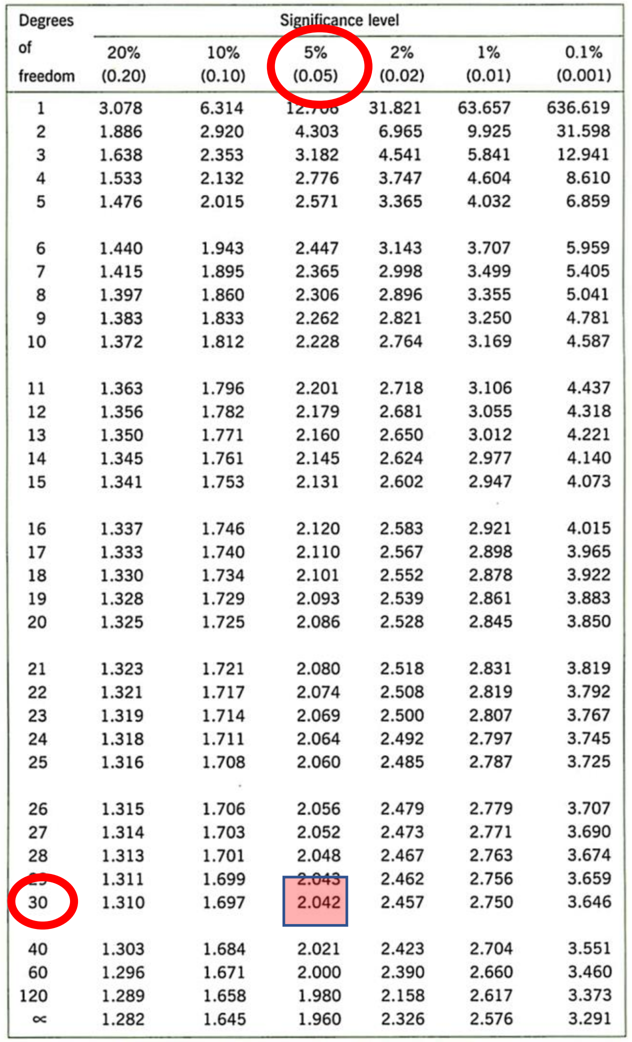
- Gosset worked at the Guinness brewery in Dublin
- In 1908 William Sealy Gosset, an Englishman publishing under the pseudonym Student, developed the t-test and t distribution.
- Gosset wanted to answer the question “What is the sampling distribution of the mean of a sample, drawn from a larger population?”
- Gosset found that existing statistical techniques using large samples were not useful for the small sample sizes that he encountered in his work.
- He started out with a resampling experiment—drawing random samples of 4 from a data set of 3,000 measurements of criminals’ height and left-middle-finger length

Use case: Student’s t-test for assessing the statistical significance of the difference between two sample means

- We want to estimate the mean $\boldsymbol{\mu}$ of a population with a normal distribution and unknown standard deviation $\boldsymbol{\sigma}$
- Our sample of size $\boldsymbol{n}$ has a mean $\bar{\boldsymbol{x}}$ and a standard deviation $\boldsymbol{S}$
- Sample size $\boldsymbol{n}$ is too small for the Central Limit Theorem to apply
- H0 : μ = m   H1 : $\mu \neq$ m
- $\boldsymbol{\alpha}$ = risk of rejecting H0 while it is correct

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim Z \longrightarrow \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

- *n = 31*
- $\frac{\bar{x}-\mu}{s/\sqrt{n}} = t_{31-1} = 3.09$ → *t = 2.042 from t table* → *3.09 > 2.042 We reject H0*
- *α = 0.05 (5%)*



Degrees of freedom	Significance level					
	20% (0.20)	10% (0.10)	5% (0.05)	2% (0.02)	1% (0.01)	0.1% (0.001)
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.158	2.617	3.373
∝	1.282	1.645	1.960	2.326	2.576	3.291

Use case: The construction of confidence intervals for the difference between two population means

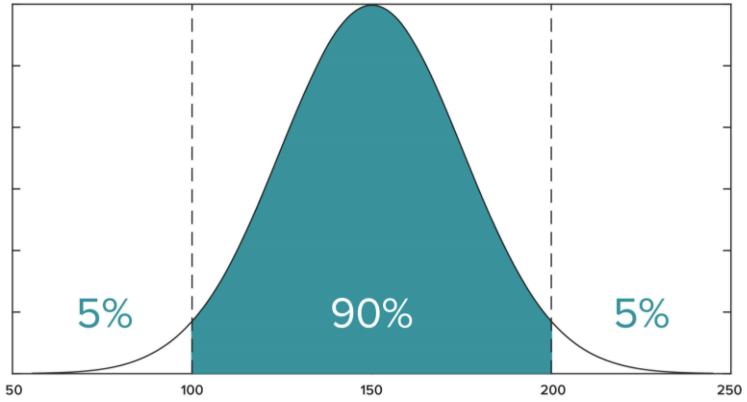
A number of different statistics can be compared, after standardization, to the t distribution, to estimate confidence intervals in light of sampling variation.

Consider a sample of size *n* for which the sample mean *x* has been calculated.

If *s* is the sample standard deviation, a 90% confidence interval around the sample mean is given by:

$$\bar{x} \pm t_{n-1}(0.05) \cdot \frac{s}{\sqrt{n}}$$

where $\mathbf{t_{n-1}}$(0 . 05) is the value of the t-statistic, with (*n* – 1) degrees of freedom that "chops off " 5% of the t-distribution at either end.



Key ideas

- The t-distribution is actually a family of distributions resembling the normal distribution but with thicker tails.
- The t-distribution is widely used as a reference basis for the distribution of sample means, differences between two sample means, regression parameters, and more.

Binomial distribution

- The binomial distribution is the frequency distribution of the number of successes (*x*) in a given number of trials (*n*) with specified probability (*p*) of success in each trial.
- There is a family of binomial distributions, depending on the values of *n* and *p*.

$$\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

- The binomial distribution would answer a question like:

If the probability of a click converting to a sale is 0.02, what is the probability of observing 0 sales in 200 clicks?

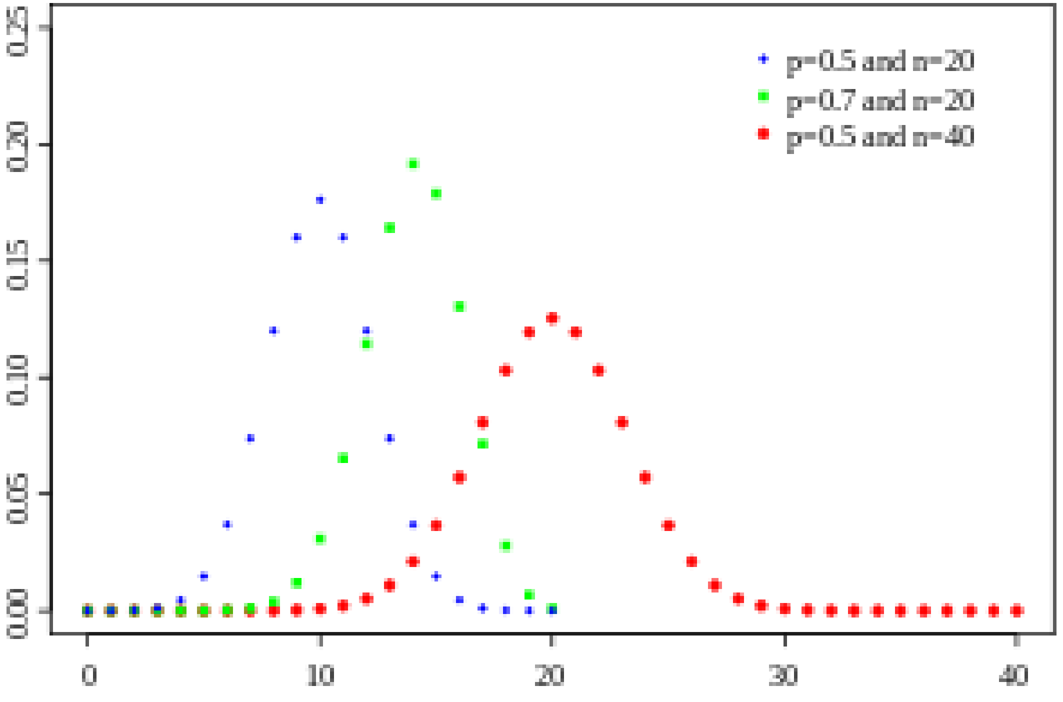
Mean:
The mean of a binomial distribution is *n* × *p*; you can also think of this as the expected number of successes in *n* trials, for success probability = *p*.

Variance:
The variance is *n* × *p (*1 − *p)* .

The binomial distribution is virtually indistinguishable from the normal distribution

- Binomial distribution is a discrete probability distribution
- Normal distribution is continuous

Calculating binomial probabilities with large sample sizes is computationally demanding, and most statistical procedures use the normal distribution,
with mean and variance, as an approximation.



Key ideas

- Binomial outcomes are important to model, since they represent, among other things, fundamental decisions (buy or don’t buy, click or don’t click, survive or die, etc.).
- A binomial trial is an experiment with two possible outcomes: one with probability *p* and the other with probability *1 – p*.
- With large *n*, and provided *p* is not too close to 0 or 1, the binomial distribution can be approximated by the normal distribution.

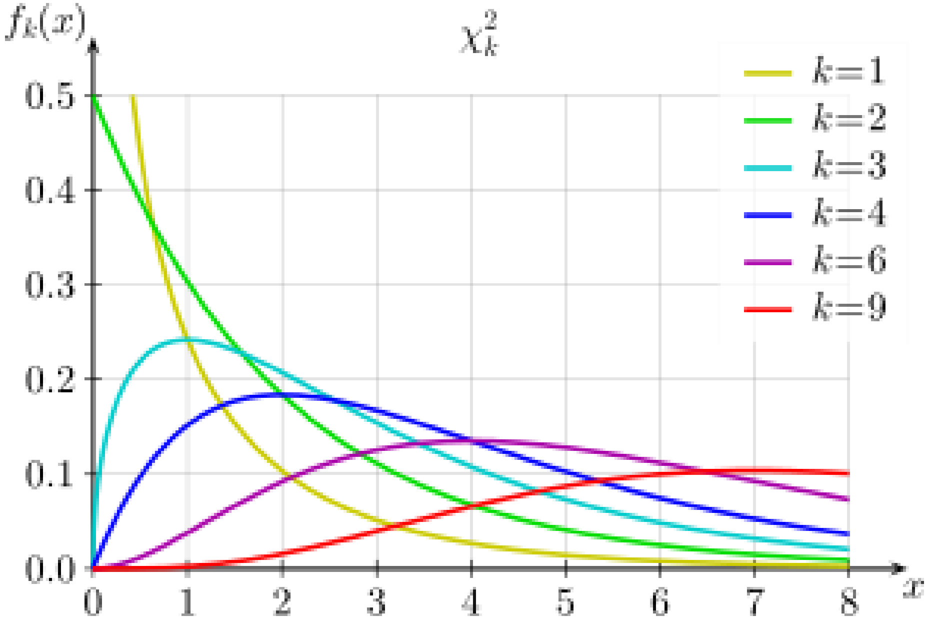
Chi-square distribution

- The chi-squared distribution (also chi-square or χ2-distribution) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.
- If $Z_1, \ldots, Z_k$ are independent standard normal random variables, then the sum of their squares,

$$Q = \sum_1^k {Z_i}^2,$$

is distributed according to the chi-squared distribution with k degrees of freedom. This is usually denoted as:

$$Q \sim {\chi_k}^2 \quad \text{or} \quad Q \sim \chi^2(k)$$



- The chi-squared distribution is used primarily in hypothesis testing, and to a lesser extent for confidence intervals for population variance when the underlying distribution is normal.

Chi-square distribution

- Chi-square statistic is a measure of the extent to which a set of observed values “fits” a specified distribution (a “goodness-of-fit” test). It is useful for determining whether multiple treatments (an “A/B/C… test”) differ from one another in their effects.

- The chi-square distribution is the distribution of this statistic under repeated resampled draws from the null model and the chi-square formula for a data table.
- A low chi-square value for a set of counts indicates that they closely follow the expected distribution

Chi-square distribution

- For example, you might want to test whether one variable (say, a row variable representing gender) is independent of another (say, a column variable representing was promoted in job”), and you have counts of the number in each of the cells of the data table.
- The statistic that measures the extent to which results depart from the null expectation of independence is the chi-square statistic.

The null hypothesis: Gender and promotion have independent distributions

Contingency table

Promoted/ Gender	Female	Male	Total
Yes	15	20	35
No	7	9	16
Total	22	29	51

$$E = \frac{\text{row total} \times \text{column total}}{\text{sample size}}$$

Promoted/ Gender	Female	Male	Total
Yes	(22*35)/51 = 15.09	19.9	35
No	6.9	9.09	16
Total	22	29	51

The chi-square statistic:

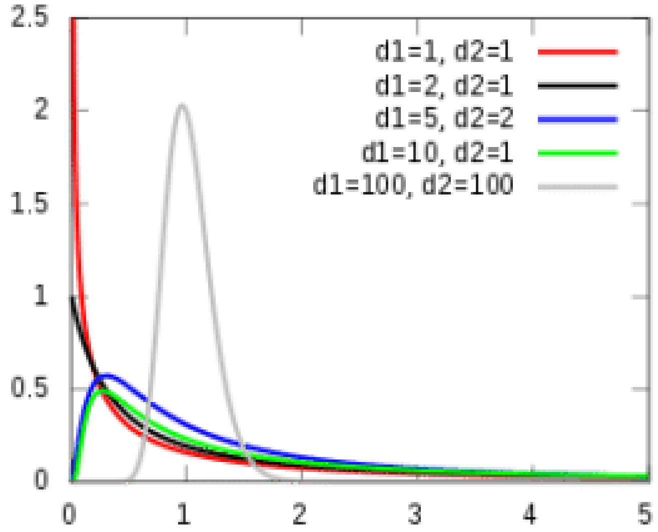
$$X^2 = \frac{(15.09\ - 15)^2}{15.09} + \frac{(19.9 - 20)^2}{19.9} + \frac{(6.9 - 7)^2}{6.9} + \frac{(9.09 - 9)^2}{9.09} = 0.003$$

Key Ideas

- The chi-square distribution is typically concerned with **counts** of subjects or items falling into categories.
- The chi-square statistic measures the extent of departure from what you would expect in a null model.

F-Distribution

- The F-distribution is a skewed distribution of probabilities similar to a chi-squared distribution. But where the chi-squared distribution deals with the degree of freedom with one set of variables, the F-distribution deals with multiple levels of events having different degrees of freedom.
- The F-distribution compares how much variance there is in the groups to how much variance there is between the groups.
- If the null hypothesis is true, then the variances would be about equal, though we use an F-table of critical values in a similar way to a t-test to determine if the values are similar enough.



$$f(x) = \frac{\Gamma\left(\frac{df_1 + df_2}{2}\right)\left(\frac{df_1}{df_2}\right)^{\frac{df_1}{2}} x^{\frac{df_1}{2} - 1}}{\Gamma\left(\frac{df_1}{2}\right)\Gamma\left(\frac{df_2}{2}\right)\left(1 + \frac{df_1}{df_2}x\right)^{\frac{df_1 + df_2}{2}}}$$

F-Distribution is used in ANOVA test

ANOVA

Analysis of variance, more commonly called ANOVA, is a statistical method that is designed to compare means of different samples. Essentially, it is a way to compare how different samples in an experiment differ from one another if they differ at all. It is similar to a t-test except that ANOVA is generally used to compare more than two samples.

(H_0) of ANOVA is that there is no difference among group means

Let’s consider that you are testing a new drug for heart disease called X. In this case you want to determine the significant effects of different dosages. So, you set up trials of 0 mg, 50 mg, and 100 mg of X in three randomly selected groups of 30 each. This is a case for ANOVA, which utilizes the F-distribution.

This compares the variance *within* a group (all the 100 mg participants for example) to the variance *between* the groups (comparing the three groups). When you run this equation, you get an F-score.

$$F = \frac{\frac{df_1 \cdot s_1^2}{\sigma_1^2} / df_1}{\frac{df_2 \cdot s_2^2}{\sigma_2^2} / df_2}$$

$$df_1 = 30 - 1 = 29$$

$$df_2 = (30 * 2) - 1 = 59$$

Key Ideas

- The F-distribution is used with experiments and linear models involving measured data.
- The F-statistic compares variation due to factors of interest to overall variation.

Poisson and related distributions

Many processes produce events randomly at a given overall rate:

events spread over space

- imperfections in a square meter of fabric
- typos per 100 lines of code

events spread over time

- visitors arriving at a website
- cars arriving at a toll plaza

Key Terms for Poisson and Related Distributions

Lambda
The rate (per unit of time or space) at which events occur.

Poisson distribution
The frequency distribution of the number of events in sampled units of time or space.

Exponential distribution
The frequency distribution of the time or distance from one event to the next event.

Weibull distribution
A generalized version of the exponential distribution in which the event rate is allowed to shift over time.

Poisson Distributions

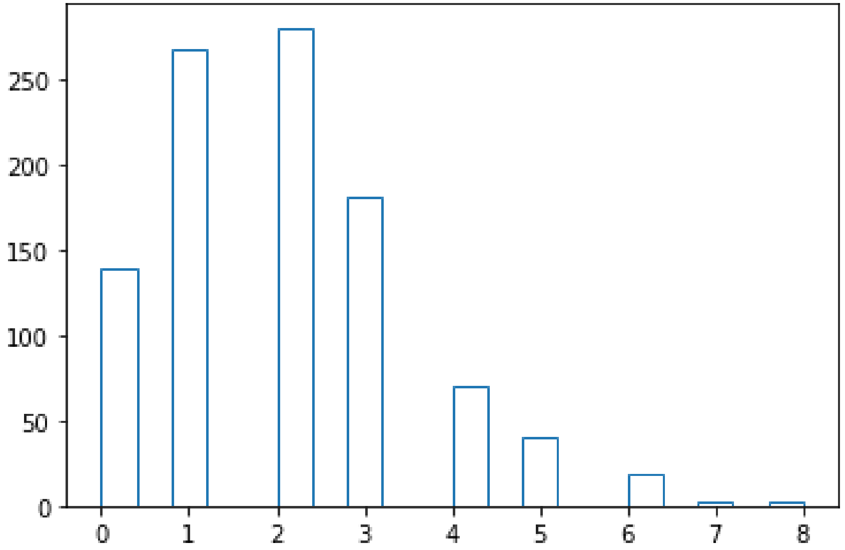
- A Poisson distribution, named after French mathematician **Siméon Denis Poisson**, can be used to estimate how many times an event is likely to occur within "X" periods of time.
- Poisson distributions are used when the variable of interest is a discrete count variable.
- λ is the mean number of events that occurs in a specified interval of time or space. The variance for a Poisson distribution is also λ

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Mean = Variance = λ

For example, if incoming customer service calls average two per minute, we can simulate 100 minutes, returning the number of calls in each of those 100 minutes.

```
stats.poisson.rvs(2, size=100)
```



Exponential Distribution

Using the same parameter λ that we used in the Poisson distribution, we can also model the distribution of the time between events: time between visits to a website or between cars arriving at a toll plaza.

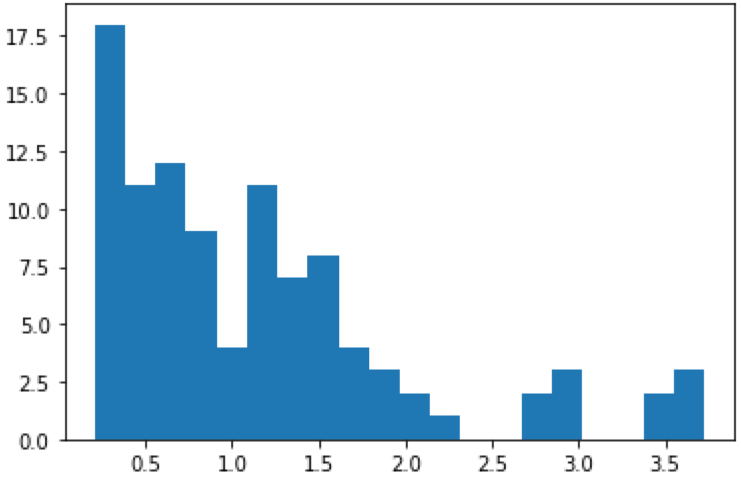
$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

The exponential distribution is the probability distribution of the time between events in a Poisson point process

Mean = $\frac{1}{\lambda}$ **Variance =** $\frac{1}{\lambda^2}$

This code below would generate 100 random numbers from an exponential distribution where the mean number of events per time period is 0.2. So, you could use it to simulate 100 intervals, in minutes, between service calls, where the average rate of incoming calls is 0.2 per minute.

```
stats.expon.rvs(0.2, size=100)
```



Estimating the Failure Rate

- In many applications, the event rate, λ, is known or can be estimated from prior data. However, for rare events, this is not necessarily so.
- Aircraft engine failure, for example, is sufficiently rare (thankfully) that, for a given engine type, there may be little data on which to base an estimate of time between failures.
- With no data at all, there is little basis on which to estimate an event rate
- However, we can make some guesses: if no events have been seen after 20 hours, you can be pretty sure that the rate is not 1 per hour.
- If there is some data but not enough to provide a precise, reliable estimate of the rate, a goodness-of-fit test (Chi squared test) can be applied to various rates to determine how well they fit the observed data.

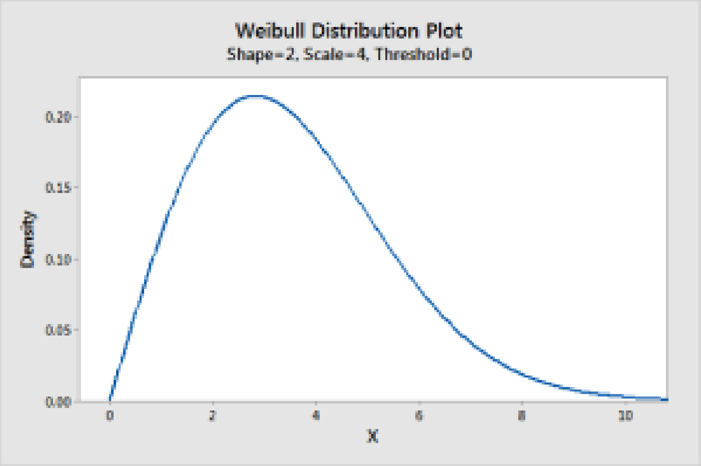
Weibull Distribution

In many cases, the event rate (lambda) does not remain constant over time.

If the period over which it changes is much longer than the typical interval between events, there is no problem; you just subdivide the analysis into the segments where rates are relatively constant

If, however, the event rate changes over the time of the interval, the exponential (or Poisson) distributions are no longer useful.

Example case: Mechanical failure: the risk of failure increases as time goes by.



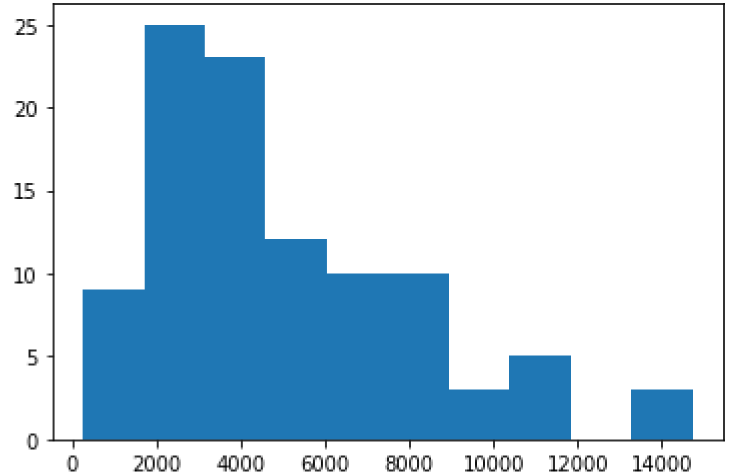
Weibull Distribution

- The *Weibull* distribution is an extension of the exponential distribution in which the event rate is allowed to change, as specified by a *shape parameter*, β. If $\beta > 1$, the probability of an event increases over time; if $\beta < 1$, the probability decreases.
- Because the Weibull distribution is used with time-to-failure analysis instead of event rate, the second parameter is expressed in terms of characteristic life, rather than in terms of the rate of events per interval.
- The symbol used is η, the Greek letter eta. It is also called the *scale* parameter.

- With the Weibull, the estimation task now includes estimation of both parameters, β and η. Software is used to model the data and yield an estimate of the best-fitting Weibull distribution.

For example, the following code would generate 100 random numbers (lifetimes) from a Weibull distribution with shape of 1.5 and characteristic life of 5,000:

```
stats.weibull_min.rvs(1.5, scale=5000, size=100)
```

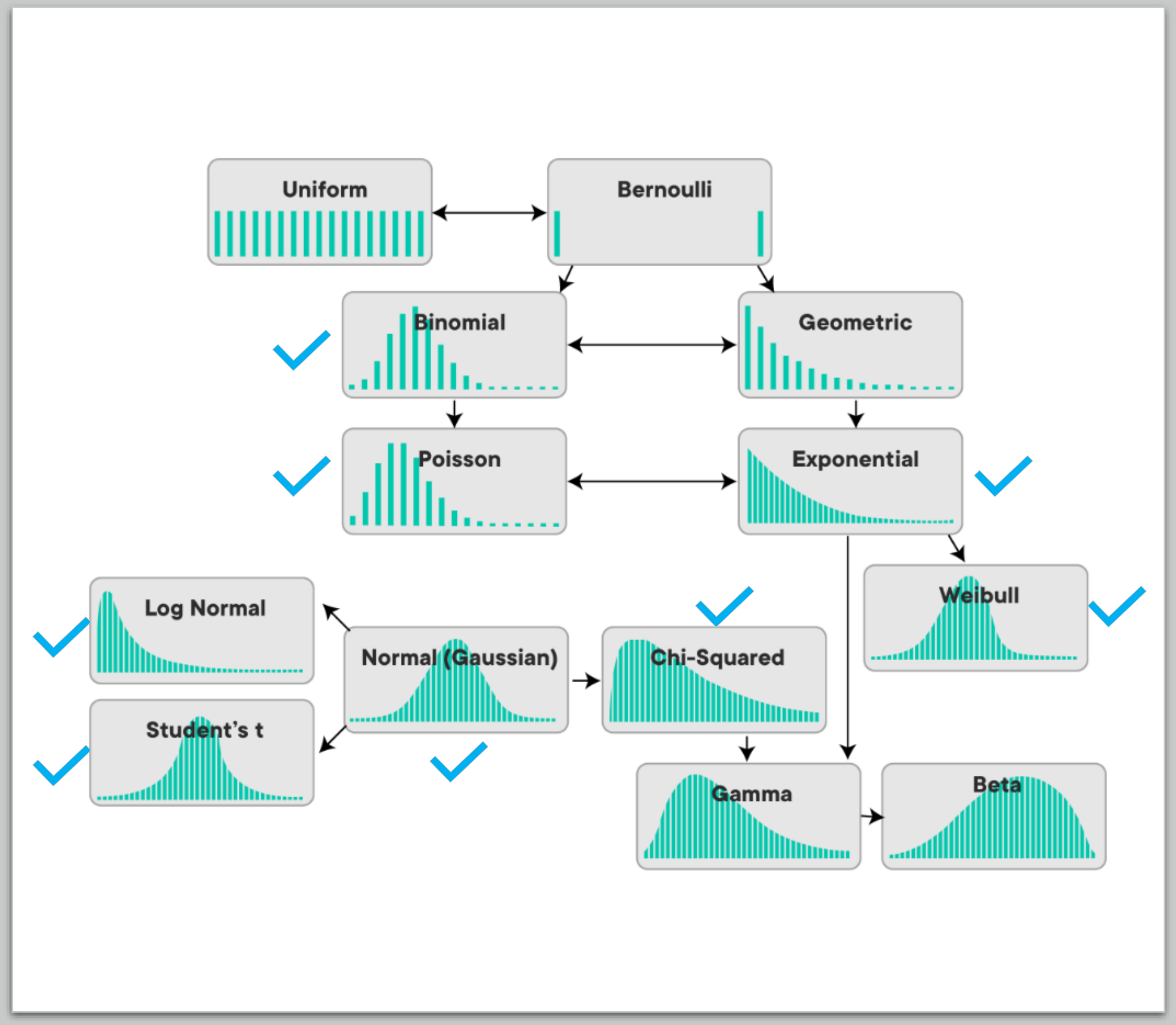


Key Ideas

- For events that occur at a constant rate, the number of events per unit of time or space can be modeled as a Poisson distribution.
- You can also model the time or distance between one event and the next as an exponential distribution.
- A changing event rate over time (e.g., an increasing probability of device failure) can be modeled with the Weibull distribution.

Summary

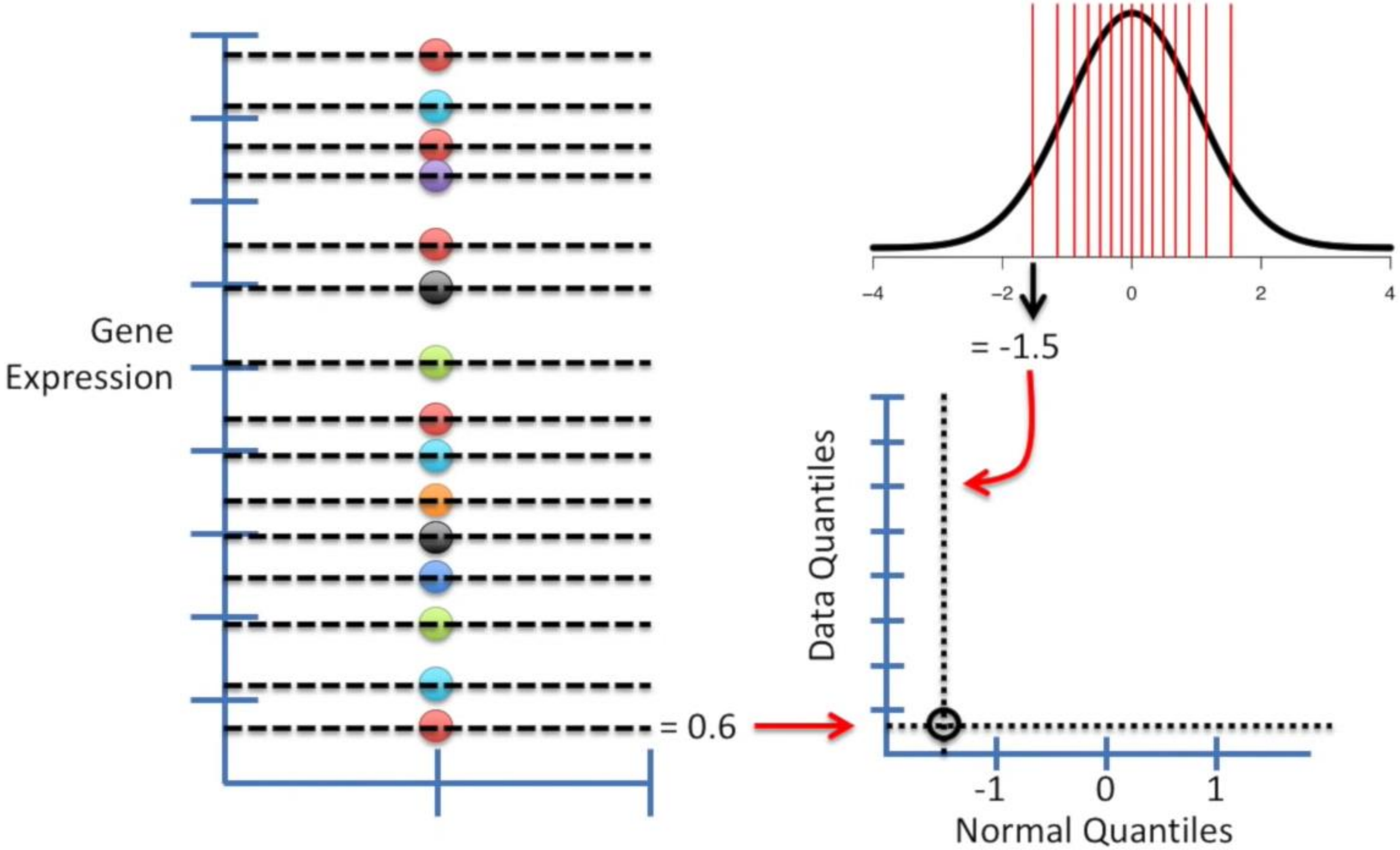
- In the era of big data, the principles of random sampling remain important when accurate estimates are needed.
- Random selection of data can reduce bias and yield a higher quality data set than would result from just using the conveniently available data.
- Knowledge of various sampling and data-generating distributions allows us to quantify potential errors in an estimate that might be due to random variation.
- At the same time, the bootstrap (sampling with replacement from an observed data set) is an attractive “one size fits all” method to determine possible error in sample estimates.



Resources

- https://statisticsbyjim.com/hypothesis-testing/test-statistic/
- https://www.youtube.com/watch?v=UetYS3PaHIo
- https://www.youtube.com/watch?v=2QeDRsxSF9M
- https://www.investopedia.com/terms/p/poisson-distribution.asp
- https://statisticsbyjim.com/anova/f-tests-anova/
- https://statisticsbyjim.com/hypothesis-testing/identify-distribution-data/
- https://magoosh.com/statistics/analysis-variance-explained/
- https://magoosh.com/statistics/f-distribution-explained/
- https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/8-chi-squared-tests
- https://www.jmp.com/en_us/statistics-knowledge-portal/t-test/one-sample-t-test.html
- https://www.youtube.com/watch?v=okjYjClSjOg

QQ Plot



Gene Expression
= 0.6
−4
−2
0
2
4
= -1.5
Data Quantiles
-1
0
1
Normal Quantiles