I only received project writeup from groups 1, 2 and 4

Group 3: please, send it today

Descriptive statistics:

Point estimation:

Point Estimation

- A sample was collected: $X_1, X_2, \ldots, X_n$
- We suspect that sample was drawn from a random variable distribution $f(x)$
- $f(x)$ has k parameters that we do not know
- Point estimates are estimates of the parameters of the $f(x)$ describing the population based on the sample
  - For exponential PDF: $f(x) = \lambda \exp(-\lambda x)$ one wants to estimate λ
  - For Bernoulli PDF: $p^x(1-p)^{1-x}$ one wants to estimate p
  - For normal PDF one wants to estimates both μ and σ
- Point estimates are uncertain: therefore, we can talk of averages and standard deviations of point estimators

Point Estimator

A point estimate of some parameter θ describing population random variable is a single numerical value $\hat{\theta}$ depending on all values $x_1, x_2, \dots x_n$ in the sample.
The sample statistic (whis a random variable $\widehat{\Theta}$ defined by a function $\widehat{\Theta}(\mathrm{X}_1, \mathrm{X}_2, \dots X_n)$) is called the
point estimator.

- There could be multiple choices for the point estimator of a parameter.
- To estimate the mean of a population, we could choose the:
  - Sample mean
  - Sample median
  - Peak of the histogram
  - ½ of (largest + smallest) observations of the sample.
- We need to develop criteria to compare estimates using statistical properties.

Unbiased Estimators Defined

The point estimator $\widehat{\Theta}$ is an unbiased estimator
for the parameter θ if:

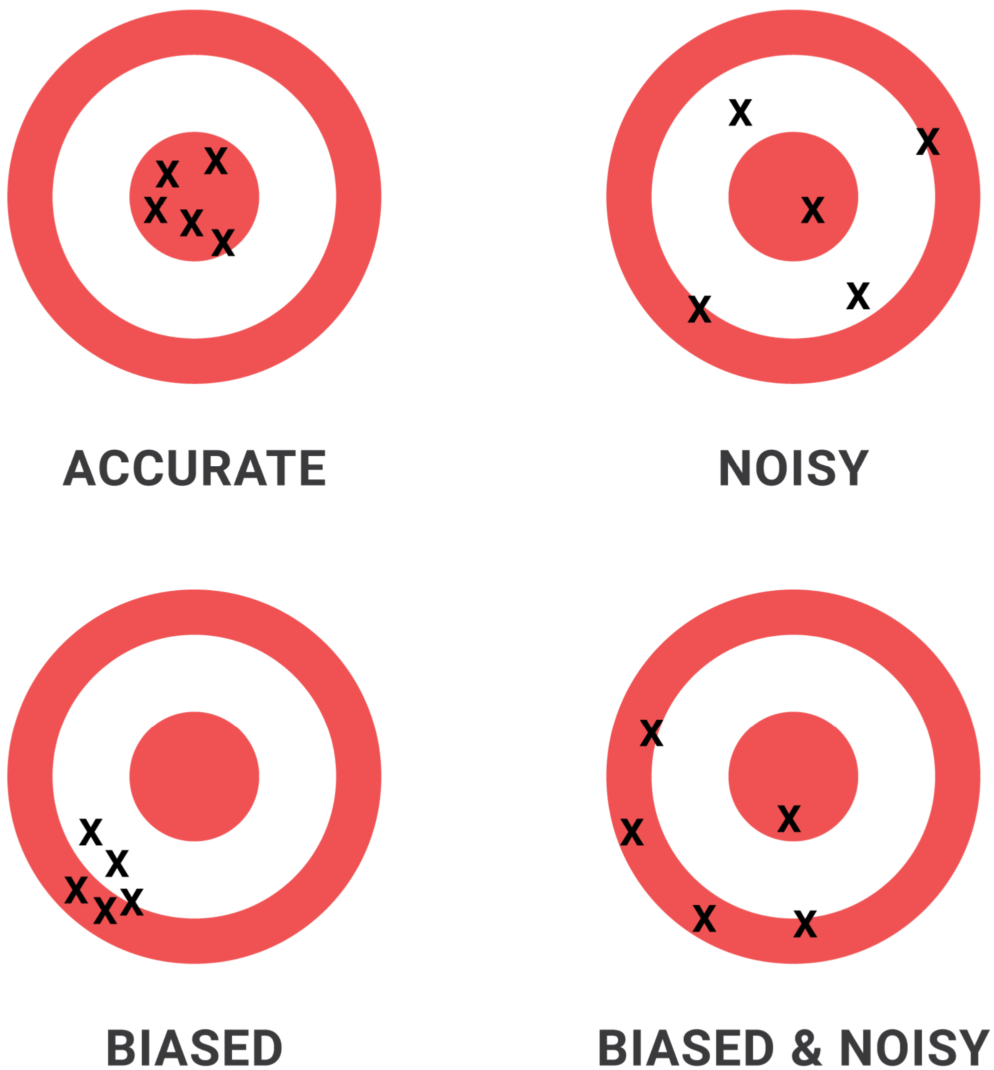
$$E(\widehat{\Theta}) = \theta \qquad (7-5)$$

If the estimator is not unbiased, then the difference:

$$E(\widehat{\Theta}) - \theta \qquad (7-6)$$

is called the bias of the estimator $\widehat{\Theta}$.

Bias vs Noise



NOISE: Daniel Kahneman, Olivier Sabony, and Cass R. Sunstein

Mean Squared Error

The **mean squared error** of an estimator $\widehat{\Theta}$ of the parameter θ is defined as:

$$\mathrm{MSE}(\widehat{\Theta}) = E(\widehat{\Theta} - \theta)^2 \qquad (7-7)$$

Can be rewritten as

$$= E[\widehat{\Theta} - E(\widehat{\Theta})]^2 + [\theta - E(\widehat{\Theta})]^2$$

$$= V(\widehat{\Theta}) + (\mathrm{bias})^2$$

Statistic #1: Sample Mean

If the values of n observations in a random sample are denoted by $x_1, x_2, \ldots, x_n$, the sample mean is

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (6-1)$$

New random variable $\overline{X}$ is a linear combination of n independent identically distributed variables $X_1, X_2, \ldots, X_n$

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Sample mean $\bar{x}$ is drawn from a random variable

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$E(\bar{X}) = \frac{n \cdot E(X_i)}{n} = \frac{n \cdot \mu}{n} = \mu$$

Sample mean, $\bar{X}$, is an <u>unbiased</u> estimator of the population mean, μ

Sample variance S^2 – is an estimator of the population variance σ^2

Sample Variance

If n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the sample variance is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad (6\text{–}3)$$

If one knows the population average, μ, one divides by n to estimate the variance

$$s(\mu)^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Why divide by n-1 instead of n?

- The sample mean $\bar{x}$ is on average closer to points $x_1, x_2, \ldots x_n$ than the true mean μ
  $\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq \sum_{i=1}^{n}(x_i - \mu)^2$
- Consider a sample of size n=1.
  Then $\bar{x} = x_1$ while $\mu \neq x_1$. Dividing by n gives $s^2 = 0$, while dividing by n-1 leaves s^2 undefined (0/0)
- For n=2, $\bar{x}$ is exactly halfway between x_1 and x_2 making its sum of squares smaller than that of μ
- Dividing by n-1 on average corrects for a smaller sum of squares: S^2 is an unbiased estimator of σ^2

Show that S^2 is unbiased estimate of σ^2

$$E(S^2) = E\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}\right) = \frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i^2 + \bar{X}^2 - 2\bar{X}X_i)\right] =$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n} X_i^2 + n\bar{X}^2 - 2\bar{X}n\bar{X}\right] =$$

$$= \frac{1}{n-1} E\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = \frac{1}{n-1}\left(nE(X_i^2) - nE(\bar{X}^2)\right)$$

$$= \frac{1}{n-1}\left(n(\mu^2 + \sigma^2) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right) = \frac{n-1}{n-1}\sigma^2 = \underline{\underline{\sigma^2}}$$

Example 7-4: Sample Variance S^2 is Unbiased

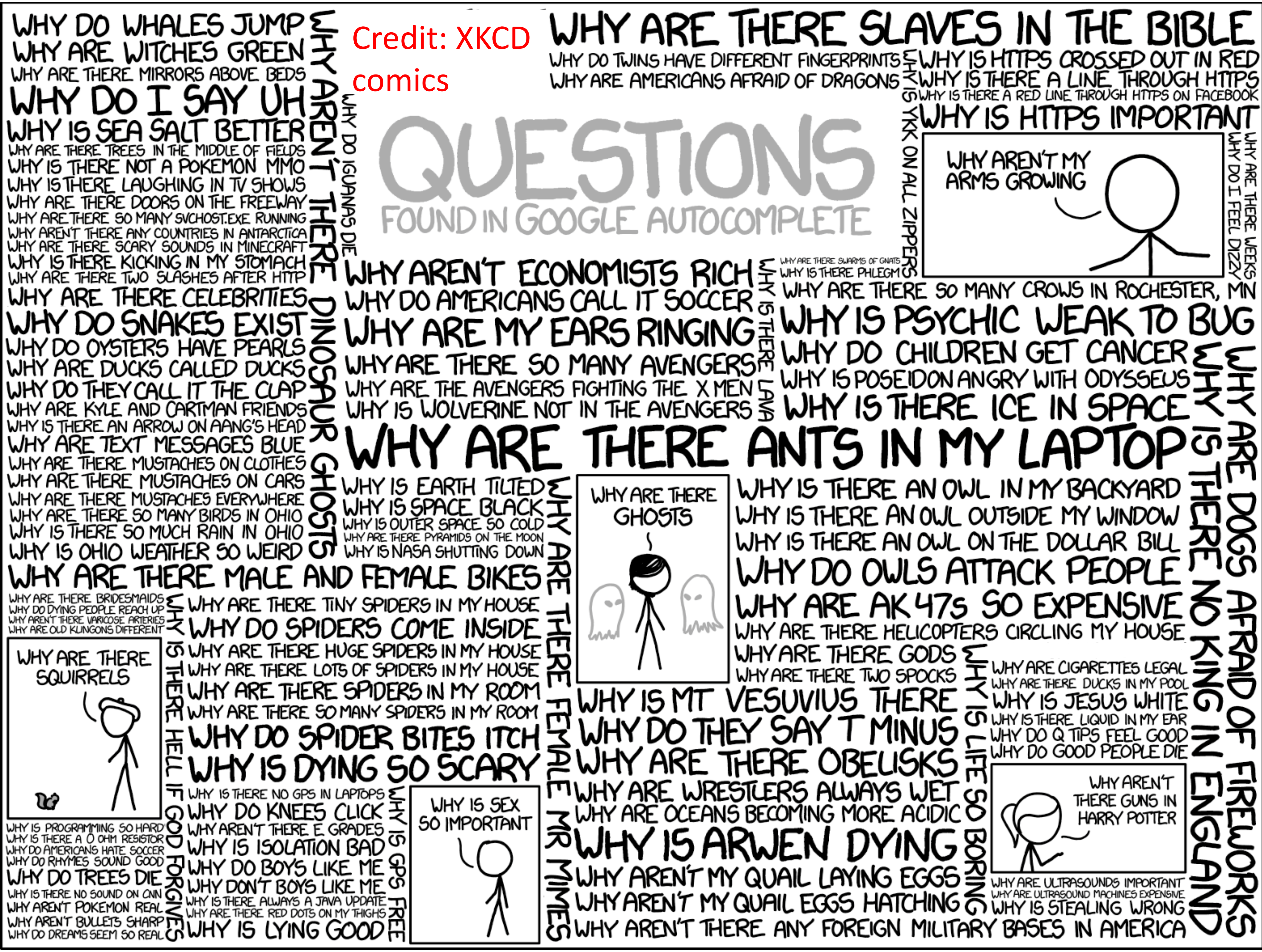
$$E\left(S^2\right) = E\left(\frac{\sum_{i=1}^{n}\left(X-\bar{X}\right)^2}{n-1}\right)$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i^2+\bar{X}^2-2\bar{X}X_i\right)\right]$$

$$= \frac{1}{n-1}\left[E\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\mu^2+\sigma^2\right)-n\left(\mu^2+\frac{\sigma^2}{n}\right)\right]$$

$$= \frac{1}{n-1}\left[n\mu^2+n\sigma^2-n\mu^2-\sigma^2\right] = \frac{1}{n-1}\left[(n-1)\sigma^2\right]$$



Credit: XKCD comics

Methods of Point Estimation

- We will cover two popular methodologies to create point estimates of a population parameter.
  - Method of moments
  - Method of maximum likelihood
- Each approach can be used to create estimators with varying degrees of biasedness and relative MSE efficiencies.

Method of moments for point estimation

What are moments?

- The p-th population moment of a random variable is the expected value of X^p
  - First moment: $\mu = \int_{\infty}^{+\infty} xf(x)\,dx$
  - Second moment: $\mu^2 + \sigma^2 = \int_{\infty}^{+\infty} x^2 f(x)\,dx$
  - p-th moment: $\int_{\infty}^{+\infty} x^p f(x)\,dx$
  - The population moment relates to the entire population

- A sample moment is calculated like its population moments but for a finite sample
  - Sample first moment = sample mean = $\frac{1}{n}\sum_{i=1}^{n} x_i$
  - Sample p-th moment $\frac{1}{n}\sum_{i=1}^{n} x_i^p$

Moment Estimators

Let $X_1, X_2, \ldots, X_n$ be a random sample from either a probability mass function or a probability density function with p unknown parameters $\theta_1, \theta_2, \ldots, \theta_p$.

The moment estimators $\hat{\Theta}_1\ \hat{\Theta}_2 \ldots, \hat{\Theta}_p$ are found by equating the first p population moments to the first p sample moments and solving the resulting simultaneous equations for the unknown parameters.

Exponential Distribution: Moment Estimator-1st moment

- Suppose that $x_1, x_2, \ldots, x_n$ is a random sample from an exponential distribution $f(x)=\lambda \exp(-\lambda x)$ with parameter λ.
- There is only one parameter to estimate, so equating population and sample first moments, we have one equation: $E(X) = \bar{x}$.
- $E(X) = 1/\lambda$ thus
  $\lambda = 1/\bar{x}$ is the 1st moment estimator.

Matlab exercise

- Generate 100,000 exponentially distributed random numbers with <u>$\lambda=3$</u>: $f(x)=\lambda \exp(-\lambda x)$
  - Use random('Exponential'…) but read the manual to know how to introduce parameters.
- Get a moment estimate of lambda based on the 1st moment
- Get a moment estimate of lambda based on the 2nd moment
  - Second moment of the exponential distribution is $E(X^2) = E(X)^2+\mathrm{Var}(X)= 1/\lambda^2 + 1/\lambda^2 = 2/\lambda^2$
- Get a moment estimate of lambda based on the 20th moment
  - Generally, p-th moment of the exponential distribution is $E(X^P) = p!/\lambda^P$

How I solved it

- Stats=100000;
- Y=random('Exponential', 1/3, Stats, 1);

%parametrization in MATLAB is 1/lambda

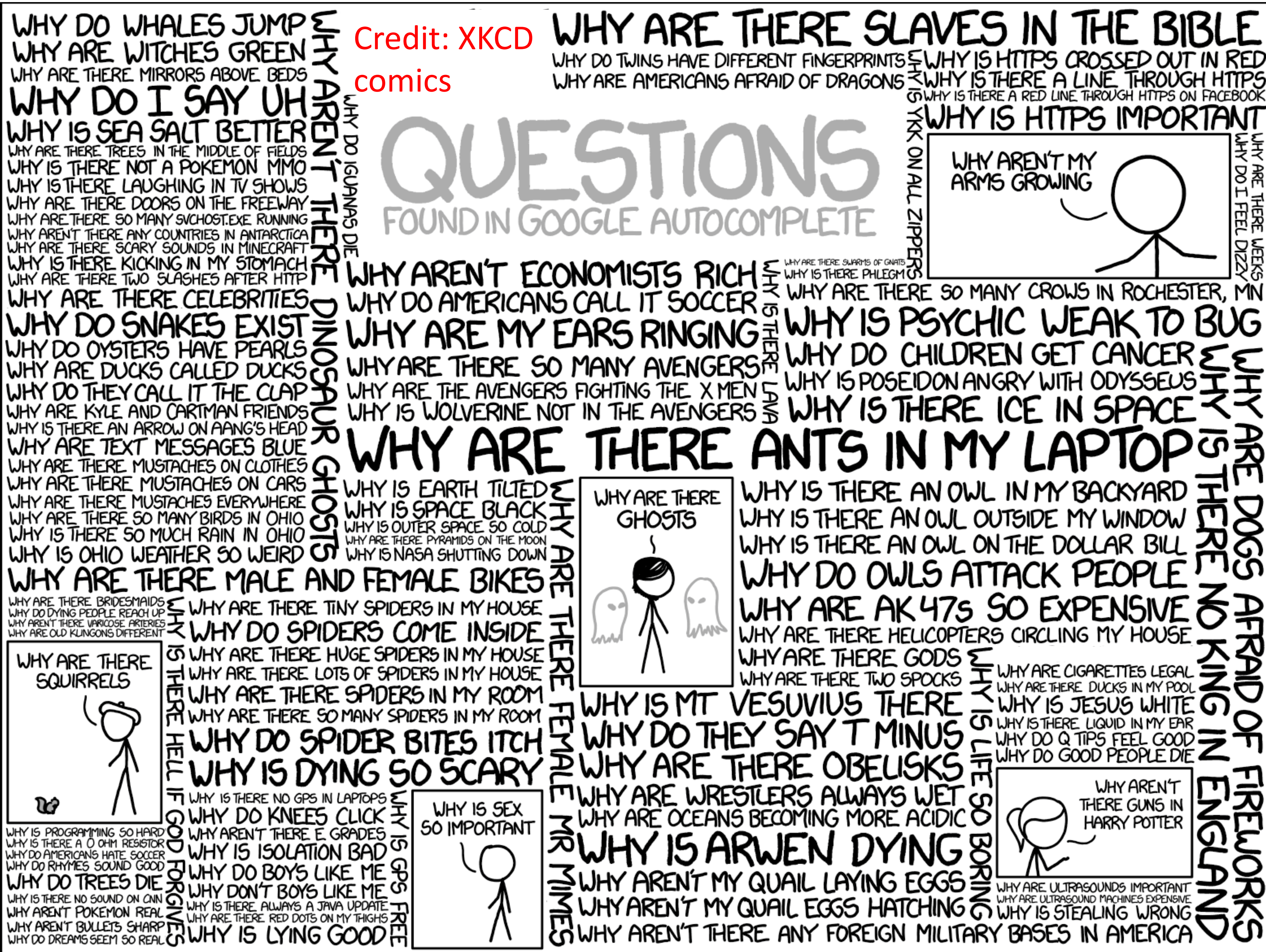
- 1/mean(Y) %matching the first moment

% ans = 3.0086

- sqrt(2/mean(Y.^2)) %matching the second moment

% ans = 3.0081

- (factorial(20)/mean(Y.^20))^(1./20) %matching the 20th moment



Credit: XKCD comics

Method of Maximum Likelihood for point estimation

Maximum Likelihood Estimators

- Suppose that X is a random variable with probability distribution $f(x, \theta)$, where θ is a single unknown parameter. Let $x_1, x_2, \ldots, x_n$ be the observed values in a random sample of size n. Then the likelihood function of the sample is the probability to get it in a random variable with PDF $f(x, \theta)$:

$$L(\theta) = f(x_1, \theta) \cdot f(x_2, \theta) \cdot \ldots \cdot f(x_n, \theta) \qquad (7\text{-}9)$$

- Note that the likelihood function is now a function of only the unknown parameter θ. The maximum likelihood estimator (MLE) of θ is the value of θ that maximizes the likelihood function $L(\theta)$.
- Usually, it is easier to work with logarithms: $l(\theta) = \ln L(\theta)$

Exponential MLE:

$$f(x_i) = \lambda e^{-\lambda x_i}$$

$$L(\lambda) = P(x_1, x_2 \ldots x_n \mid \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} =$$

$$= \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

$$\ln L(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum x_i = 0$$

$$\hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\overline{X}}$$

Same as 1^{st} moment estimator

Example 7-11: Exponential MLE

Let X be a exponential random variable with parameter λ. The likelihood function of a random sample of size n is:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

$$\ln L(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{X}} \quad \text{(same as moment estimator)}$$

Bernoulli MLE

$$f(x, p) = p^{x}(1-p)^{1-x}$$

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} =$$

$$= p^{\sum x_i}(1-p)^{n-\sum x_i}$$

$$\ln L(p) = \left(\sum x_i\right)\ln p + \left(n - \sum x_i\right)\ln(1-p)$$

$$\frac{d \ln L(p)}{dp} = \frac{\sum x_i}{p} - \frac{n-\sum x_i}{1-p} = 0 \text{ at } \hat{p}$$

$$0 = \frac{(1-\hat{p})\sum x_i - \hat{p}\left(n - \sum x_i\right)}{\hat{p}(1-\hat{p})}$$

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example 7-9: Bernoulli MLE

Let *X* be a Bernoulli random variable. The probability mass function is $f(x;p) = p^x(1-p)^{1-x}$, $x = 0, 1$ where P is the parameter to be estimated. The likelihood function of a random sample of size *n* is:

$$L(p) = p^{x_1}(1-p)^{1-x_1} \cdot p^{x_2}(1-p)^{1-x_2} \cdot \ldots \cdot p^{x_n}(1-p)^{1-x_n}$$

$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

$$\ln L(p) = \left(\sum_{i=1}^{n} x_i\right)\ln p + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-p)$$

$$\frac{d \ln L(p)}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\left(n - \sum_{i=1}^{n} x_i\right)}{(1-p)} = 0$$

$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$ (same as moment estimator)

Normal MLE for μ

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

$$L(\mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left(-\frac{\sum (x_i - \mu)^2}{2\sigma^2}\right)$$

$$\ln \mathcal{L}(\mu, \sigma) = -n \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\frac{d \ln \mathcal{L}(\mu, \sigma)}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \text{ at } \hat{\mu}$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example 7-10: Normal MLE for μ

Let X be a normal random variable with unknown mean μ and variance σ^2. The likelihood function of a random sample of size n is:

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{(2\sigma^2)}}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\ln L(\mu) = \frac{-n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{d\ln L(\mu)}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X} \text{ (same as moment estimator)}$$

Example 7-11: Normal MLE for σ^2

Let *X* be a normal random variable with the estimate of mean μ determined by MLE (see the previous slide) and an unknown variance σ^2. The likelihood function of a random sample of size n is:

$$L(\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{(2\sigma^2)}}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\ln L(\sigma) = \frac{-n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{d \ln L(\sigma)}{d\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} \quad \text{(biased estimator)}$$

MLE for Poisson distribution

$$f(x_1, \ldots, x_n|\lambda) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}$$

$$= \frac{e^{-n\lambda}\lambda^{\Sigma_1^n x_i}}{x_1! \ldots x_n!}$$

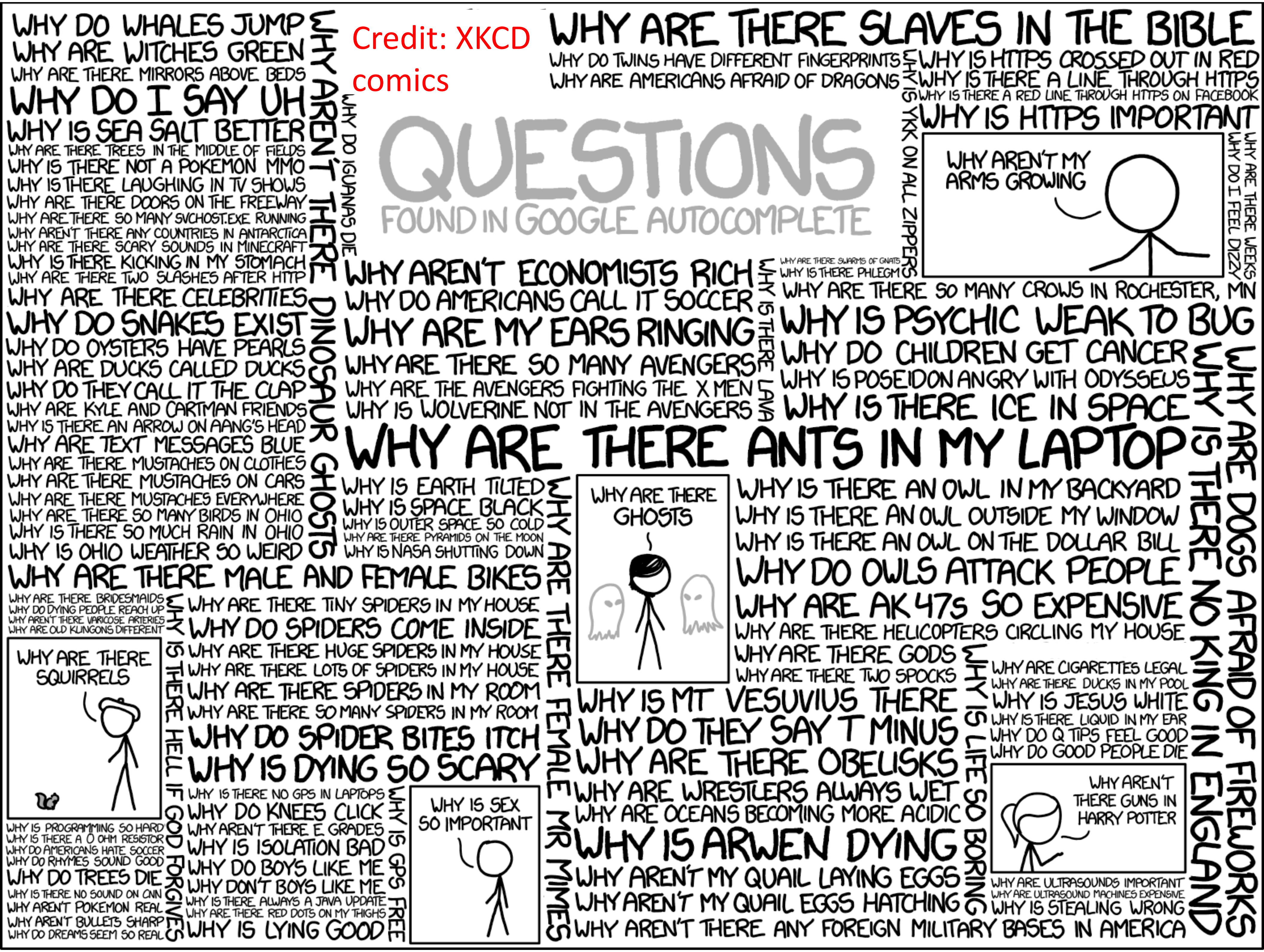
$$\log f(x_1, \ldots, x_n|\lambda) = -n\lambda + \sum_1^n x_i \log \lambda - \log c$$

where $c = \prod_{i=1}^n x_i!$ does not depend on λ, and

$$\frac{d}{d\lambda} \log f(x_1, \ldots, x_n|\lambda) = -n + \frac{\sum_1^n x_i}{\lambda}$$

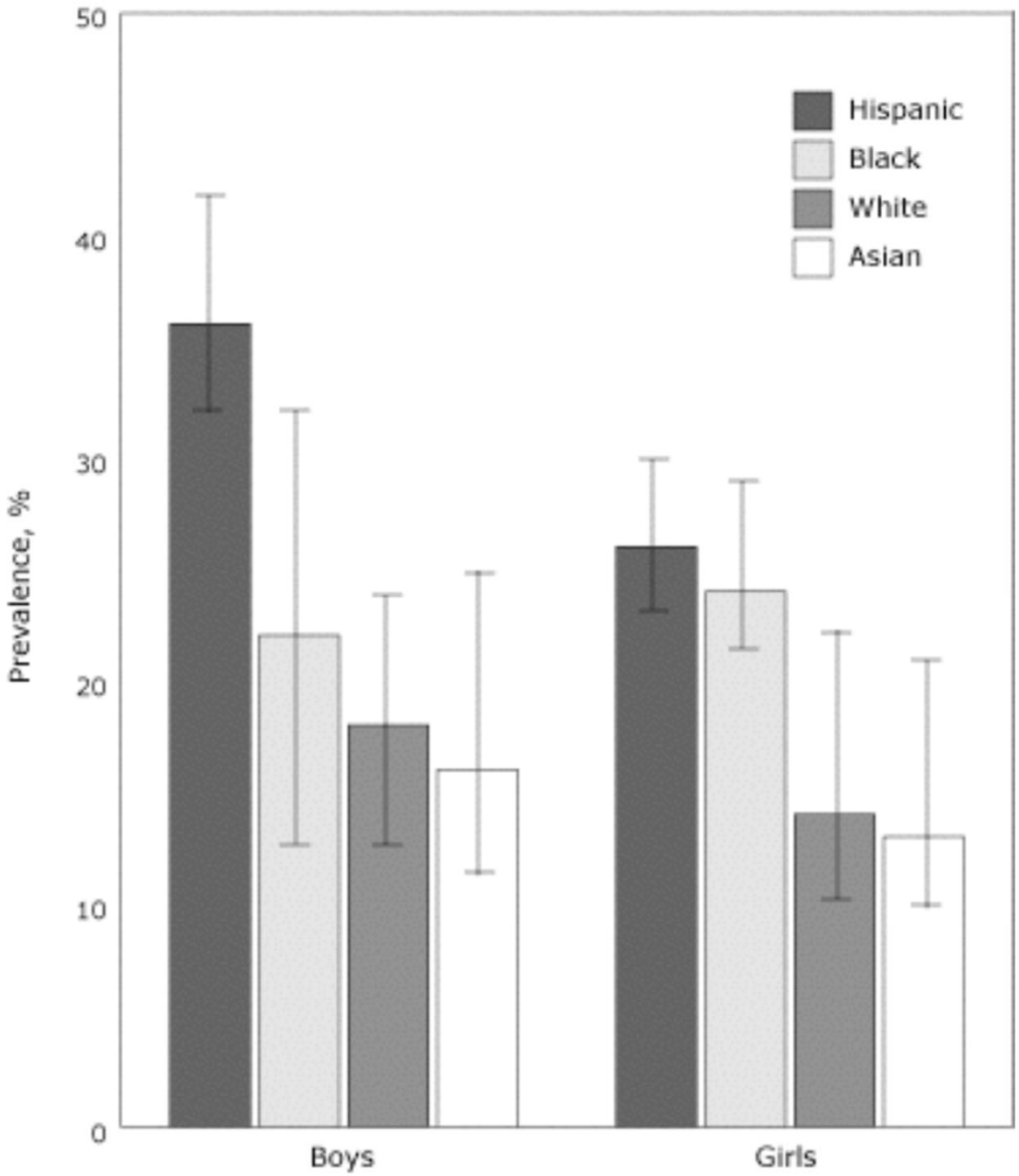
By equating to zero, we obtain that the maximum likelihood estimate $\hat{\lambda}$ equals

$$\hat{\lambda} = \frac{\sum_1^n x_i}{n}$$



Credit: XKCD comics

Confidence Intervals



Prevalence (with 95% CI bars) of obesity among New York City public elementary schoolchildren, by sex and race/ethnicity, 2003.

(source: CDC.GOV)

What do those bars actually mean?

Vol 446 | 8 March 2007 | doi:10.1038/nature05610 nature

ARTICLES

Patterns of somatic mutation in human cancer genomes

What does confidence interval mean?

The numbers of passenger and driver mutations present can be estimated from these results (see Supplementary Methods). Of the 921 base substitutions in the primary screen, 763 (95% confidence interval, 675–858) are estimated to be passenger mutations. Therefore, the large majority of mutations found through sequencing cancer genomes are not implicated in cancer development, even when the search has been targeted to the coding regions of a gene family of high candidature. However, there are an estimated 158 driver mutations (95% confidence interval, 63–246), accounting for the observed positive selection pressure. These are estimated to be distributed in 119 genes (95% confidence interval, 52–149). The number of samples containing a driver mutation is estimated to be 66 (95% confidence interval, 36–77). The results, therefore, provide statistical evidence for a large set of mutated protein kinase genes implicated in the development of about one-third of the cancers studied.

- We have talked about how a parameter can be estimated from sample data. However, it is important to understand how good is the estimate obtained.

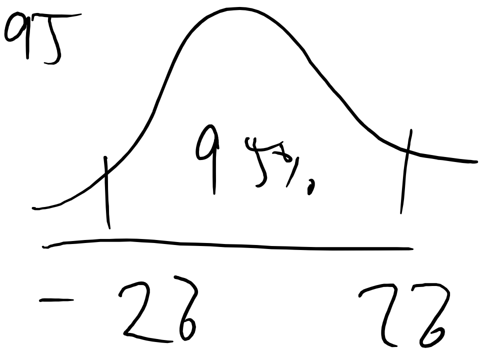
- Bounds that represent an interval of plausible values for a parameter are an example of an **interval estimate**.

Two-sided confidence intervals

- Calculated based on the sample $X_1, X_2,\ldots,X_n$
- Characterized by:
  - lower- and upper- confidence limits *L* and *R*
  - the confidence coefficient $1-\alpha$
- Objective: for two-sided confidence interval, find L and R such that
  - Prob$(\mu>R)=\alpha/2$
  - Prob$(\mu<L)=\alpha/2$
  - Therefore, Prob$(L<\mu<R)=1-\alpha$
- For one-sided confidence interval, say, upper bound of μ, find R that
  - Prob$(\mu>R)=\alpha$
- **Assume standard deviation sigma is known**

Consider $1-\alpha = 95\% = 0.95$

$\alpha = 0.05$; $\frac{\alpha}{2} = 0.025$



$z_{\alpha/2} = 1.96 \Rightarrow Prob(Z > z_{\alpha/2}) = \frac{\alpha}{2}$

$$Prob\left(-z_{\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$Prob\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

For one sided lower bound on μ

$$Prob\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha}\right) \rightarrow$$

$$\mu > \bar{X} - z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

$z_{\alpha} = 1.65 <$

$z_{\alpha/2} = 1.96$

Exercise

Ishikawa et al. (Journal of Bioscience and Bioengineering 2012) studied the force with which bacterial biofilms adhere to a solid surface.

Five measurements for a bacterial strain of Acinetobacter gave readings 2.69, 5.76, 2.67, 1.62, and 4.12 dyne-cm2.

Assume that the standard deviation is known to be 0.66 dyne-cm2

(a) Find 95% confidence interval for the mean adhesion force

(b) If scientists want the width of the confidence interval to be below 0.55 dyne-cm2 what number of samples should be?

Ishikawa et al. (Journal of Bioscience and Bioengineering 2012) studied the force with which bacterial biofilms adhere to a solid surface. Five measurements for a bacterial strain of Acinetobacter gave readings 2.69, 5.76, 2.67, 1.62, and 4.12 dyne-cm2. Assume that the standard deviation is known to be 0.66 dyne-cm2

(a) Find 95% confidence interval for the mean adhesion force

(b) If scientists want the width of the confidence interval to be below 0.55 dyne-cm2 what number of samples should be?

a) 95% CI for μ, $n = 5$ $\sigma = 0.66$ $\bar{x}$ = 3.372, $z = 1.96$

$$\bar{x} - z\sigma/\sqrt{n} \le \mu \le \bar{x} + z\sigma/\sqrt{n}$$

$$3.372 - 1.96(0.66/\sqrt{5}) \le \mu \le 3.372 + 1.96(0.66/\sqrt{5})$$

$$2.79 \le \mu \le 3.95$$

b) Width is $2z\sigma/\sqrt{n} = 0.55$, therefore n = [2zσ/0.55]2 = [2(1.96)(0.66)/0.55]2 = 22.13

Round up to $n = 23$.