HW3 has been posted.
Answers will be posted next Thursday

Projects due next Tuesday 10/24/2023 before class

One file per group (cc all group members) with all 4 assignments

Descriptive statistics:

Sample mean and its variance

Standard error vs Standard deviation

Some Definitions

- The random variables $X_1, X_2, \ldots, X_n$ are a random sample of size *n* if:
  a) The X_i are independent random variables.
  b) Every X_i has the same probability distribution.

  Such $X_1, X_2, \ldots, X_n$ are also called independent and identically distributed (or i. i. d.) random variables
- A **statistic** is any function of the observations in a random sample.
- The probability distribution of a statistic is called a **sampling distribution**.

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

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

Mean & Variance of a Linear Function

$Y = c_1X_1 + c_2X_2 + \ldots + c_pX_p$

$$E(Y) = c_1E(X_1) + c_2E(X_2) + \ldots + c_pE(X_p) \quad (5\text{-}25)$$

$$\mathrm{V}(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) + 2\sum_{i<j}\sum c_ic_j \operatorname{cov}(X_iX_j) \quad (5\text{-}26)$$

If $X_1, X_2, \ldots, X_p$ are independent, then $\operatorname{cov}(X_iX_j) = 0,$

$$V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) \quad (5\text{-}27)$$

IMPORTANT:

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

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

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

$$V(\bar{X}) = \frac{n \cdot V(X_i)}{n^2} = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\text{Stand. dev.}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ is a random sample of size n is taken from a population with mean μ and finite variance σ^2, and **any distribution**. If $\bar{X}$ is the sample mean, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad (7-1)$$

for large n, is the standard normal distribution.

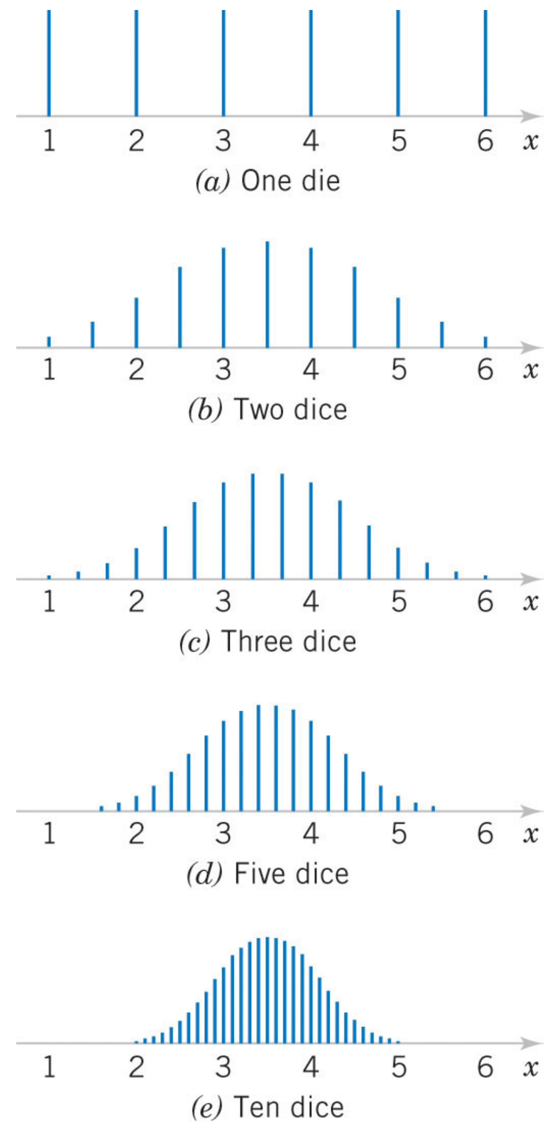
If $X_1, X_2, \ldots, X_n$ are themselves normally distributed $-$ for any n

Test CLT for your own random variable

- Go to: https://onlinestatbook.com/stat_sim/sampling_dist/
- Select "Custom" at the top and use mouse to sketch the PMF of your own random variable
- Select "mean" and n=5 in the third panel
- Choose "Animated" in the second panel and use number_of_experiments=5 to see one sample being generated
- Repeat with number_of _experiments =10,000
- Now select "mean" and n=25 in the fourth panel
- Skewness and Curtosis are measures of how good is the normal (Gaussian) fit (choose "fit normal")

Sampling Distributions of Sample Means

Figure 7-1 Distributions of average scores from throwing dice.
Mean = (6+1)/2=3.5
Sigma^2 = $[(6-1+1)^2-1]/12=2.92$
Sigma=1.71



Formulas

$$\mu = \frac{b+a}{2} = 3.5$$

$$\sigma_X^2 = \frac{(b-a+1)^2-1}{12} = 35/12$$

$$\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$$

show Matlab

Matlab exercise

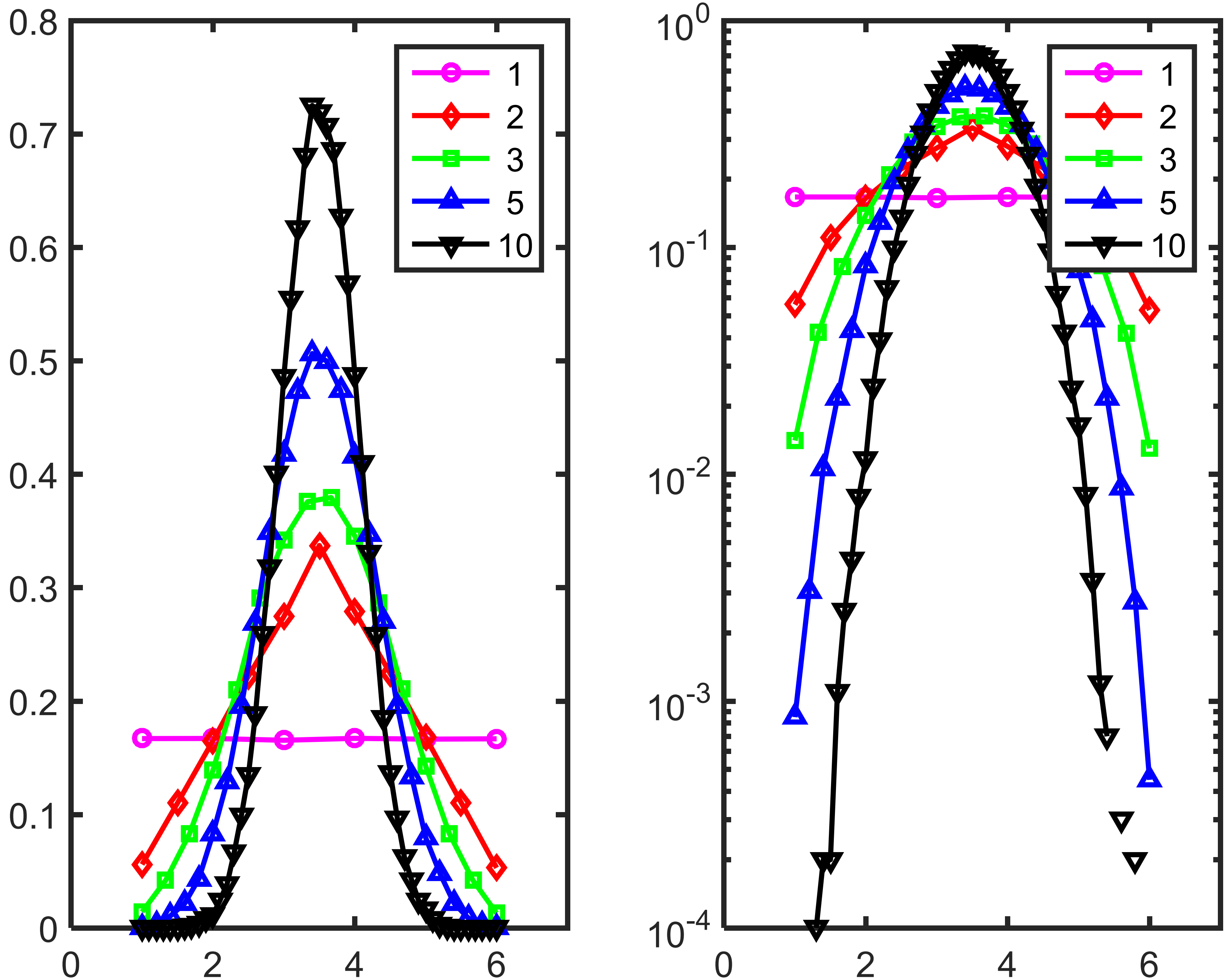
- Do a numerical experiment: generate a sample of size n by rolling n fair dice
- Calculate the sample mean $$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$
- Repeat Stats=100,000 times
- Generate PDFs of sample means for different samples sizes: n=1, n=2, n=3, n=5, and n=10
- Plot them in the same (semi-logarithmic) figure
- **What do you see?**
- Template is at the website: central_limit_theorem_template.m

How did I do it?

- **Stats=100000;**
- **figure;**
- **for n=[1,2,3,5,10];**
- **r_sample=floor(6.*rand(Stats,n))+1;**
- **sample_mean=sum(r_sample,2)./n;**
- **step=1./n;**
- **[a,b1]=hist(sample_mean,1:step:6);**
- **pdf_r1=a./sum(a)./step;**
- **semilogy(b1,pdf_r1,'o-'); hold on;**
- **end;**
- **legend('1','2','3','5','10');**

Matlab demonstration

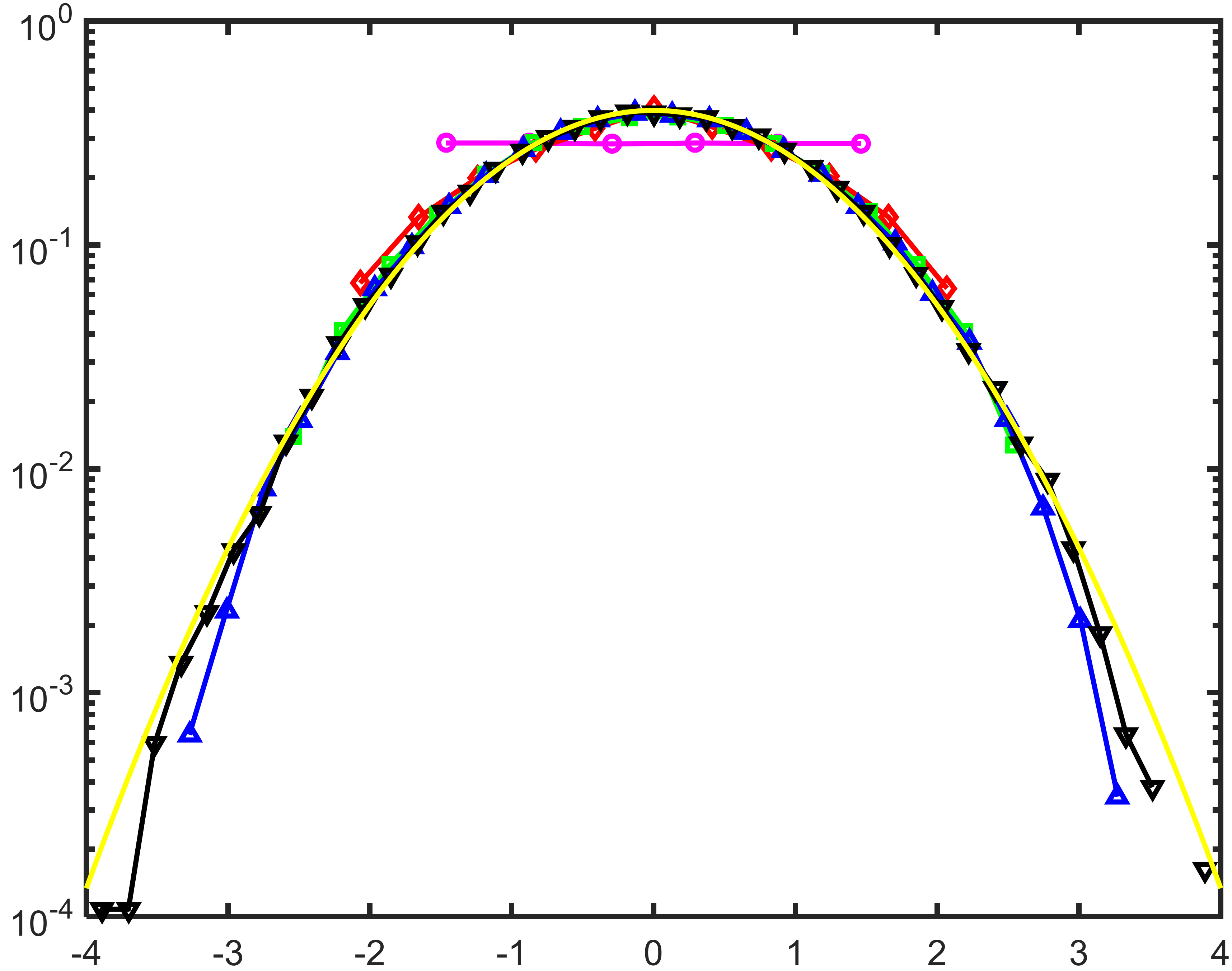
- Stats=100000; N=10;
- r_table=floor(6.*rand(Stats,N))+1;
- %%
- r1=r_table(:,1);
- step=1; [a,b1]=hist(r1,1:step:6);
- pdf_r1=a./sum(a)./step;
- figure; hold on; subplot(1,2,1); plot(b1,pdf_r1,'mo-'); hold on; axis([0 7 0 0.2]); subplot(1,2,2); semilogy(b1,pdf_r1,'mo-'); hold on; axis([0 7 1e-3 1]);
- %%
- r2=(r_table(:,1)+r_table(:,2))./2;
- step=0.5; [a,b2]=hist(r2,1:step:6); pdf_r2=a./sum(a)./step;
- subplot(1,2,1); plot(b2,pdf_r2,'rd-'); axis([0 7 0 0.4]); subplot(1,2,2); semilogy(b2,pdf_r2,'rd-');
- %%
- r3=(r_table(:,1)+r_table(:,2)+r_table(:,3))./3;
- step=1./3; [a,b3]=hist(r3,1:step:6); pdf_r3=a./sum(a)./step;
- subplot(1,2,1); plot(b3,pdf_r3,'gs-'); axis([0 7 0 0.4]); subplot(1,2,2); semilogy(b3,pdf_r3,'gs-');
- %%
- r5=sum(r_table(:,1:5),2)./5;
- step=1./5; [a,b5]=hist(r5,1:step:6); pdf_r5=a./sum(a)./step;
- subplot(1,2,1); plot(b5,pdf_r5,'b^-'); axis([0 7 0 0.6]); subplot(1,2,2); semilogy(b5,pdf_r5,'b^-'); axis([0 7 1e-4 1]);
- %%
- r10=sum(r_table(:,1:10),2)./10;
- step=1./10; [a,b10]=hist(r10,1:step:6); pdf_r10=a./sum(a)./step;
- subplot(1,2,1); plot(b10,pdf_r10,'kv-'); axis([0 7 0 0.8]); legend(num2str([1,2,3,5,10]'));
- subplot(1,2,2); semilogy(b10,pdf_r10,'kv-'); legend(num2str([1,2,3,5,10]'));



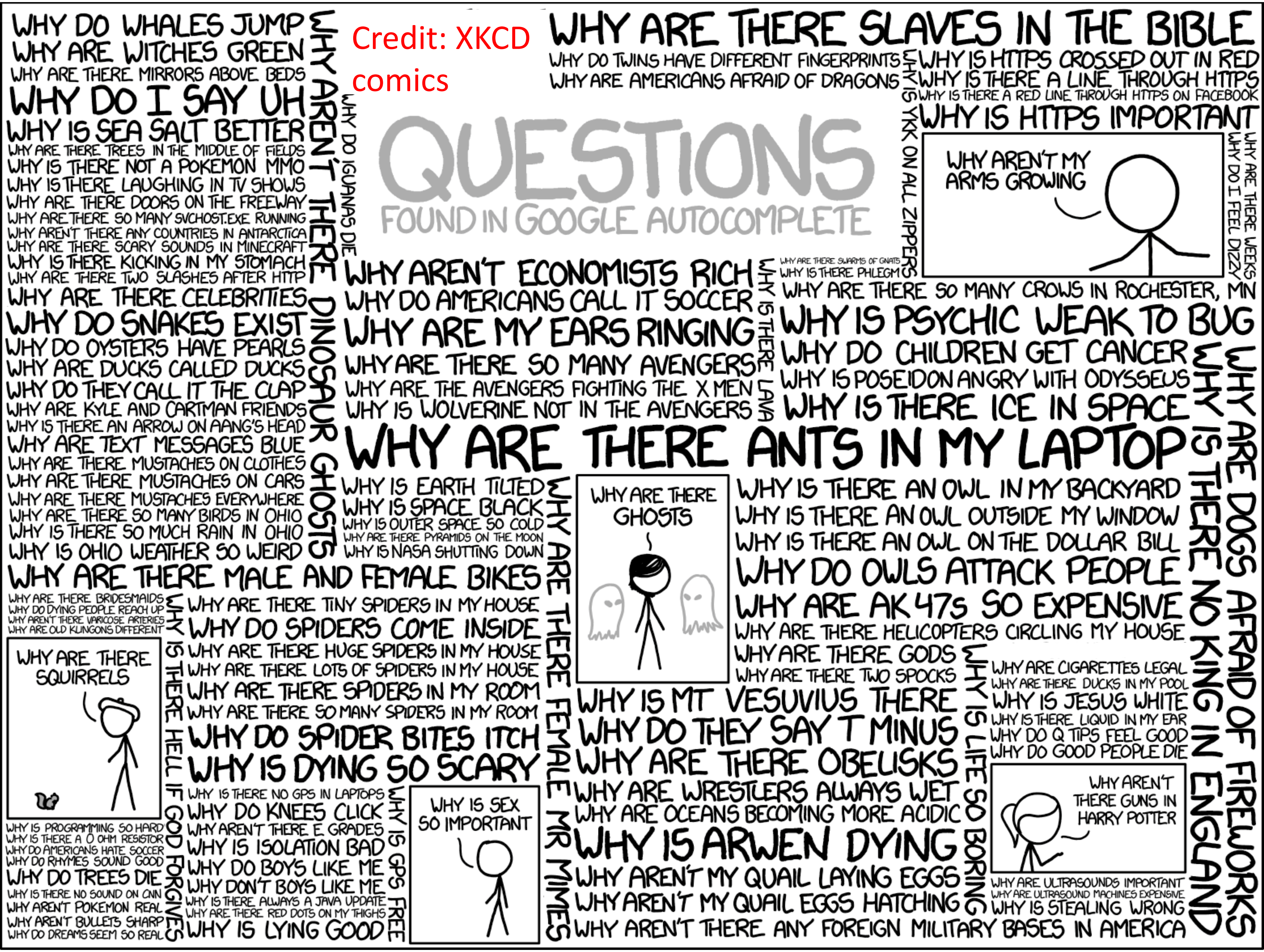
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
1
2
3
5
10
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
2
4
6
1
2
3
5
10

Matlab demonstration; part 2

- %%Now plot all of them normalized to 0 and std 1
- sigma=sqrt(35/12);
- mu=3.5;
- figure;
- sigma1=sigma;
- semilogy((b1-mu)./sigma1,pdf_r1.*sigma1,'mo-');
- axis([-4 4 1e-3 1]);
- hold on;
- %%
- sigma2=sigma./sqrt(2);
- semilogy((b2-mu)./sigma2,pdf_r2.*sigma2,'rd-');
- %%
- sigma3=sigma./sqrt(3);
- semilogy((b3-mu)./sigma3,pdf_r3.*sigma3,'gs-');
- %%
- sigma5=sigma./sqrt(5);
- semilogy((b5-mu)./sigma5,pdf_r5.*sigma5,'b^-');
- axis([-4 4 1e-4 1]);
- %%
- sigma10=sigma./sqrt(10);
- semilogy((b10-mu)./sigma10,pdf_r10.*sigma10,'kv-');
- axis([-4 4 1e-4 1]);
- %%
- %Let's see how well does the Gaussian fits it
- x=-4:0.1:4;
- semilogy(x,1./sqrt(2*pi)*exp(-x.^2./2),'y-');



10^0
10^-1
10^-2
10^-3
10^-4
-4
-3
-2
-1
0
1
2
3
4



Credit: XKCD comics

Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. What is the approximate probability that a random sample of n = 25 resistors will have an average resistance of less than 95 ohms?

Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. What is the approximate probability that a random sample of n = 25 resistors will have an average resistance of less than 95 ohms?

$\mu = 100$ ohms, $\sigma = 10$ ohms, $n = 25$

$$\mu_{\bar{X}} = \mu; \quad \sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2 \text{ ohms}$$

$$Z_{\bar{X}} = \frac{95 - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{95 - 100}{2} = -2.5$$

$$\text{Prob}(\bar{X} < 95) = \Phi(Z_{\bar{X}}) = \Phi(-2.5) =$$

$$= 0.0062$$

Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. What is the approximate probability that a random sample of *n* = 25 resistors will have an average resistance of less than 95 ohms?

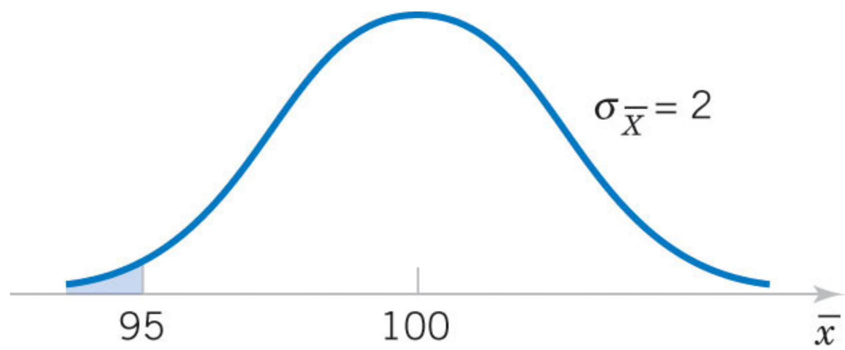


Figure 7-2 Desired probability is shaded

Answer:

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2.0$$

$$\Phi\left(\frac{\overline{X} - \mu}{\sigma_{\overline{X}}}\right) = \Phi\left(\frac{95 - 100}{2}\right)$$

$$= \Phi(-2.5) = 0.0062$$

Two Populations

We have two independent populations. What is the distribution of the difference of their sample means?

The sampling distribution of $\bar{X}_1 - \bar{X}_2$

has the following mean anad variance:

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$$

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Sampling Distribution of a Difference in Sample Means

- If we have two independent populations with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2,
- And if X-bar$_1$ and X-bar$_2$ are the sample means of two independent random samples of sizes n_1 and n_2 from these populations:
- Then the sampling distribution of:

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \qquad (7\text{-}4)$$

  is approximately standard normal, if the conditions of the central limit theorem apply.
- If the two populations are normal, then the sampling distribution is exactly standard normal.

Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a random variable with μ_{old}=5000 hours and σ_{old}=40 hours (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters to μ_{new}=5050 hours and σ_{new}=30 hours (new).

Random samples of 16 components manufactured using "old" process and 25 components using "new" process are chosen.

What is the probability new sample mean is at least 25 hours longer than old?

Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a random variable with μ_{old}=5000 hours and σ_{old}=40 hours (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters to μ_{new}=5050 hours and σ_{new}=30 hours (new).

Random samples of 16 components manufactured using "old" process and 25 components using "new" process are chosen.

What is the probability new sample mean is at least 25 hours longer than old?

$$\sigma_{\bar{X}_{old}} = \frac{\sigma_{old}}{\sqrt{16}} = 10 \text{ hrs}$$

$$\sigma_{\bar{X}_{new}} = \frac{\sigma_{new}}{\sqrt{25}} = 6 \text{ hrs}$$

$$\sigma_{TOT} = \sqrt{\sigma_{\bar{X}_{old}}^2 + \sigma_{\bar{X}_{new}}^2} = \sqrt{100+36} \approx 11.7 \text{ hrs}$$

$$\mu_{new} - \mu_{old} = 50 \text{ hrs}$$

$$z = \frac{25-(50)}{11.7} = -2.14$$

$$Prob(z > -2.14) = 0.9840$$

Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a normal-distributed random variable with parameters shown (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters mu and sigma as shown (new).

Random samples are selected from the "old" process and "new" process as shown.

What is the probability new sample mean is at least 25 hours longer than old?

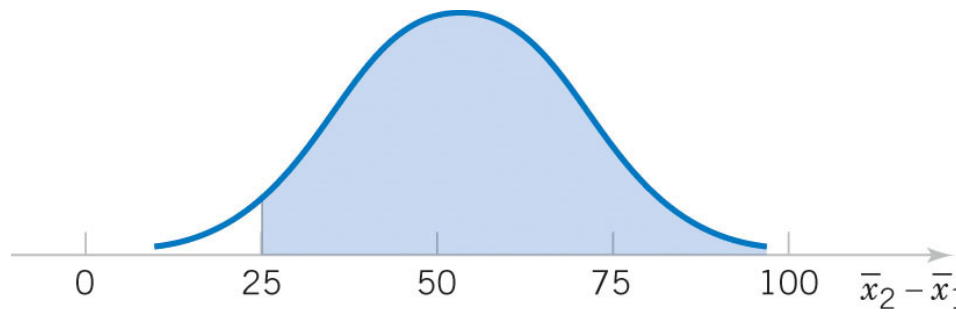
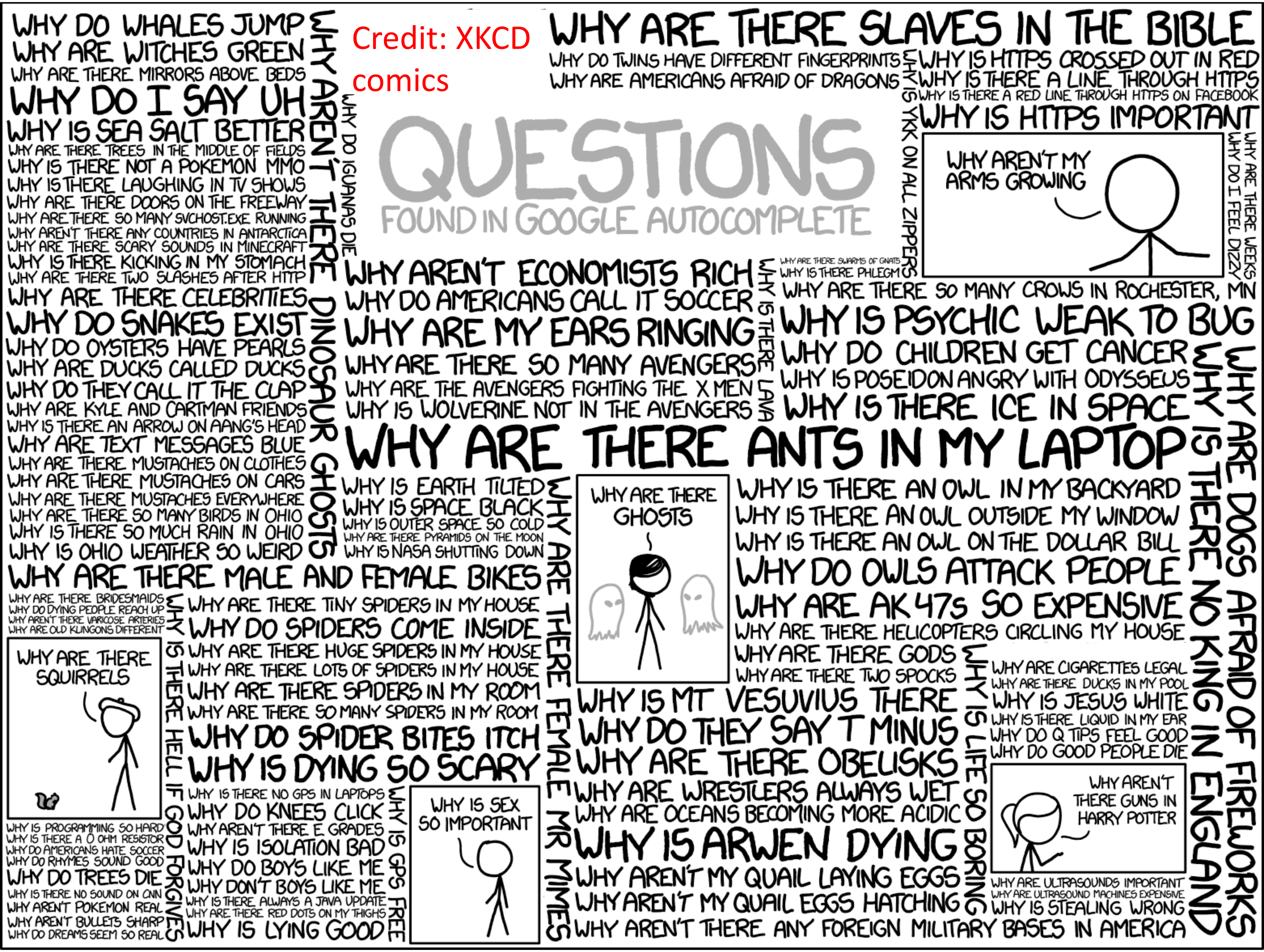


Figure 7-4 Sampling distribution of the sample mean difference.

	Process		
	Old (1)	New (2)	Diff (2-1)
mu =	5,000	5,050	50
sigma =	40	30	50
n =	16	25	
Calculations			
s / √*n* =	10	6	11.7
		z =	-2.14
		$P(xbar_2 - xbar_1 > 25) = P(Z>z) =$	0.9840



Credit: XKCD comics

Descriptive statistics:

Point estimation:

Some Definitions

- The random variables $X_1, X_2,…,X_n$ are a random sample of size n if:
  a) The X_i are independent random variables.
  b) Every X_i has the same probability distribution.

  Such $X_1, X_2,…,X_n$ are also called independent and identically distributed (or i. i. d.) random variables
- A **statistic** is any function of the observations in a random sample.
- The probability distribution of a statistic is called a **sampling distribution**.

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 estimates

Point Estimator

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

Mean Squared Error

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

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

Can be rewritten as

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

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?

- A k-th moment of a random variable is the expected value $E(X^k)$
  - First moment: $\mu = \int_{\infty}^{+\infty} xf(x)\,dx$
  - Second moment: $\mu^2 + \sigma^2 = \int_{\infty}^{+\infty} x^2 f(x)\,dx$
- A 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 k-th moment $\frac{1}{n}\sum_{i=1}^{n} x_i^k$

Moment Estimators

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

The moment estimators $\hat{\Theta}_1, \hat{\Theta}_2, ..., \hat{\Theta}_m$ are found by equating the first m population moments to the first m 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.