

The Normal Distribution, Hypothesis Testing, and t tests

- Chapter(s) in basic textbook
 - Wild & Seber (2000). Chance encounters: A first course in data analysis and inference. John Wiley & Sons.
- Howell: Chapters 3, 4, and 7
 - As well as chapters 1, 2, and 5

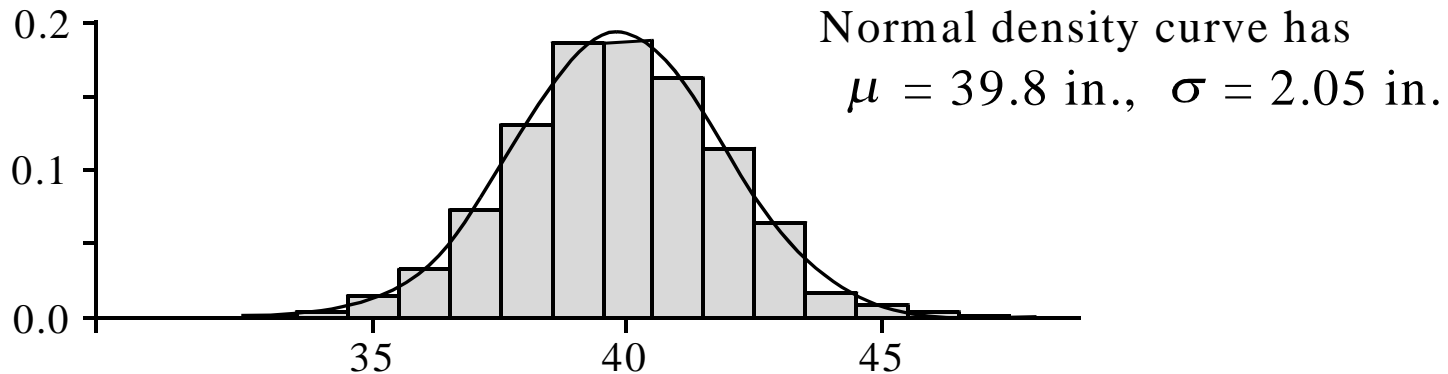
The Normal distribution

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

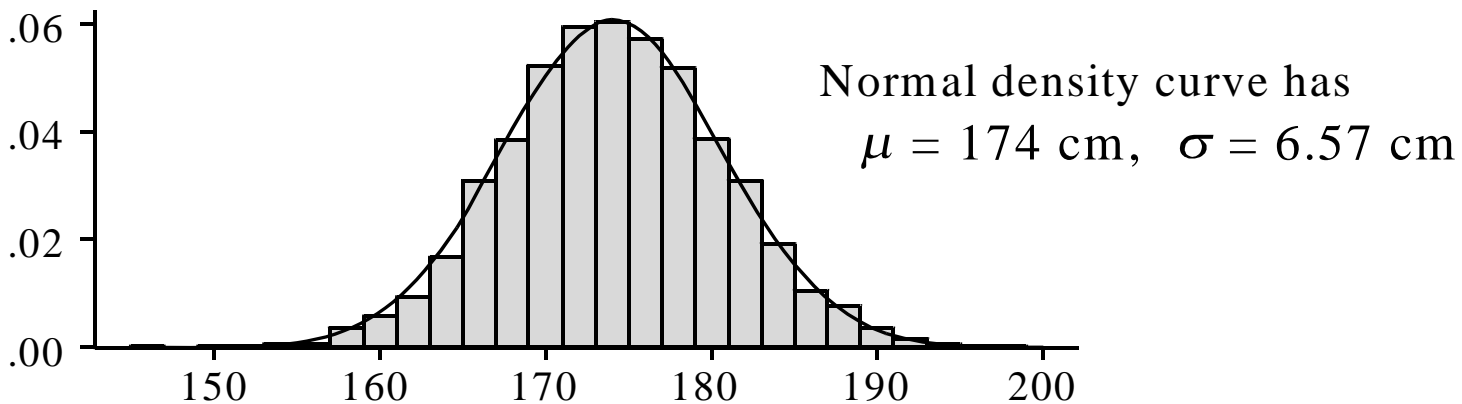


The Normal distribution

- Chest measurements of 5738 Scottish soldiers by Belgian scholar Lambert Quetelet (1796-1874)
 - First application of the Normal distribution to human data



(a) Chest measurements of Quetelet's Scottish soldiers (in.)

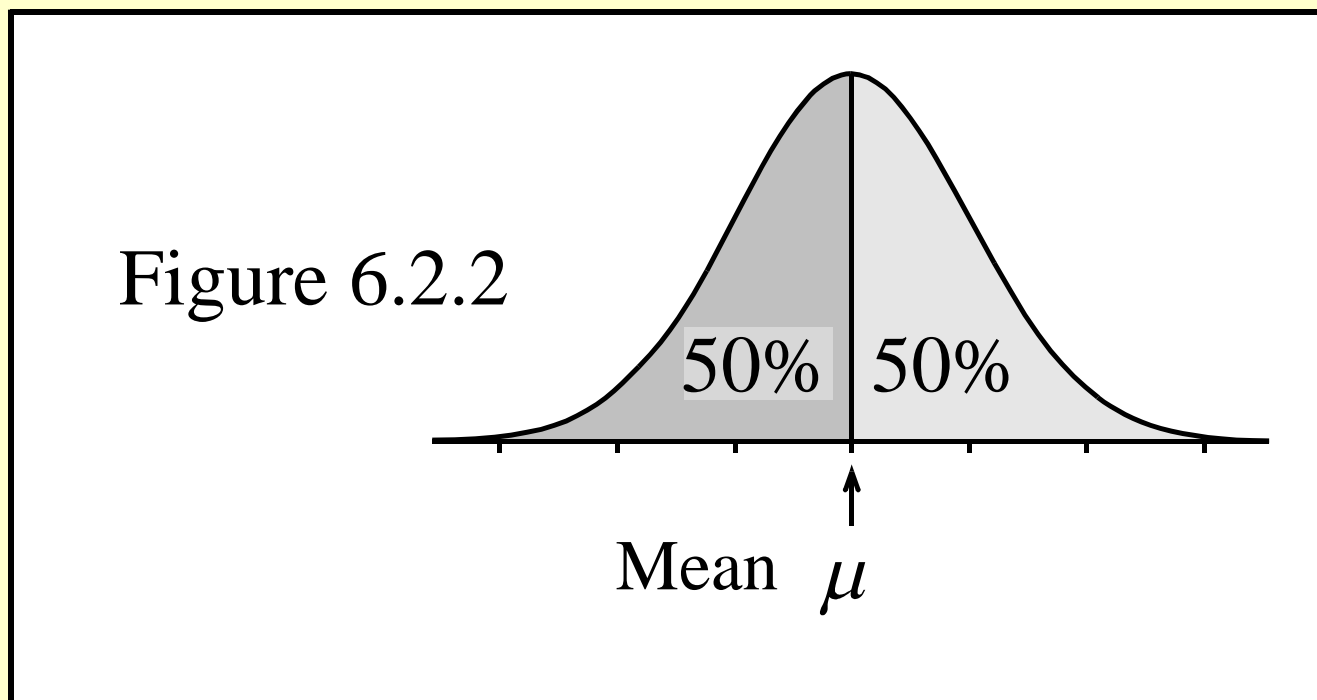


(b) Heights of the 4294 men in the workforce database (cm)

Figure 6.2.1 Two standardized histograms with approximating Normal density curves.

The Normal distribution density curve

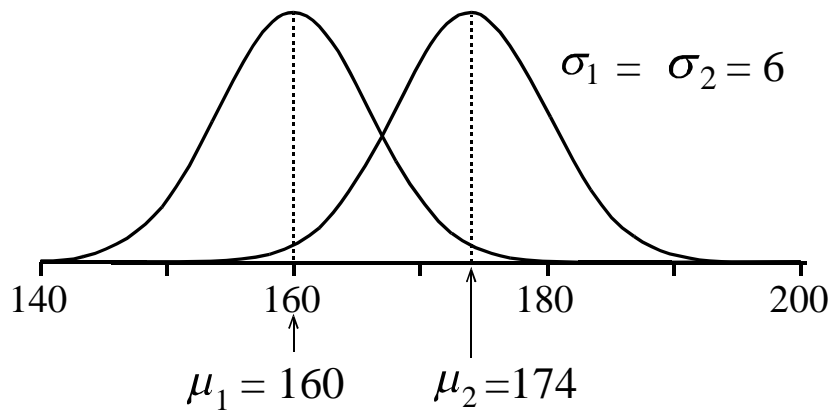
- Is symmetric about the mean
- Mean = Median



Effects of μ and σ

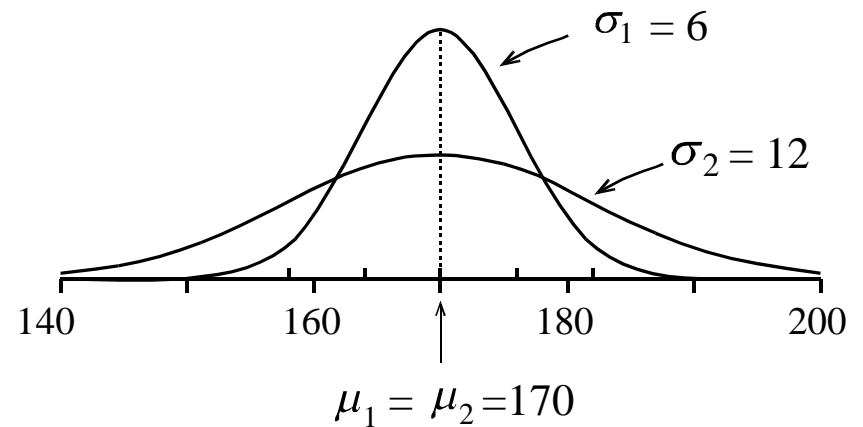
(a) Changing μ

shifts the curve along the axis



(b) Increasing σ

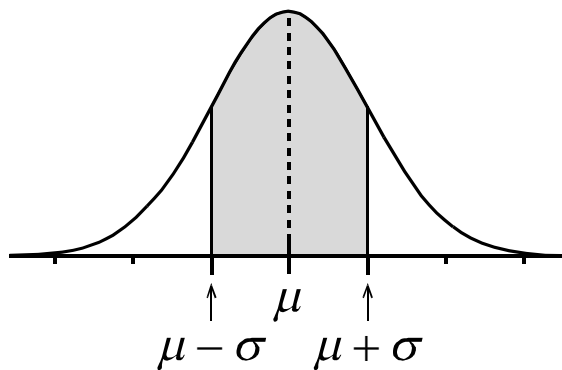
increases the spread and flattens the curve



Understanding the standard deviation σ

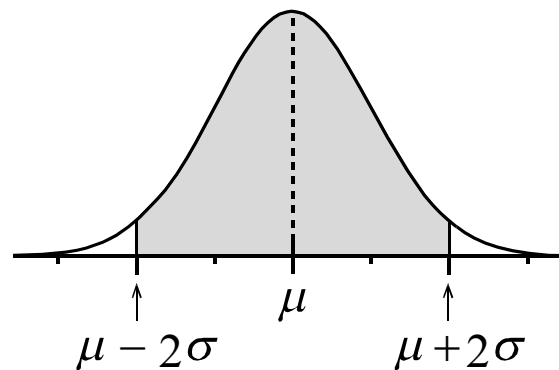
(c) Probabilities and numbers of standard deviations

Shaded area = 0.683



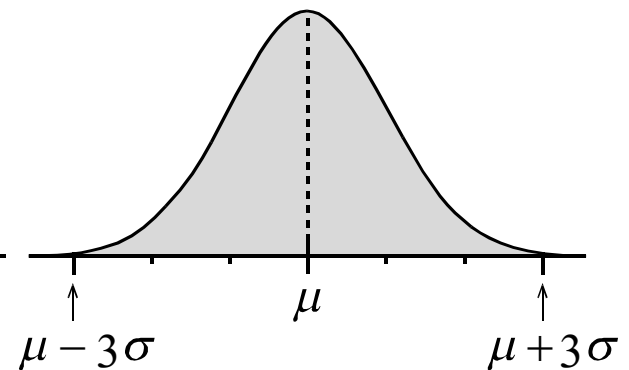
68% chance of falling
between $\mu - \sigma$ and $\mu + \sigma$

Shaded area = 0.954



95% chance of falling
between $\mu - 2\sigma$ and $\mu + 2\sigma$

Shaded area = 0.997



99.7% chance of falling
between $\mu - 3\sigma$ and $\mu + 3\sigma$

Standardizing

- For any X_i value from a Normal population with mean μ and standard deviation σ , the value

$$Z = \frac{X_i - \mu}{\sigma}$$

tells us how many standard deviations from the mean the value X_i is located

- $Z \sim N(0,1)$: standard Normal distribution
 - Z score
 - Normal deviate

The sample mean has a sampling distribution

Sampling batches of Scottish soldiers and taking chest measurements. Pop mean = 39.8 in, Pop sd = 2.05 in

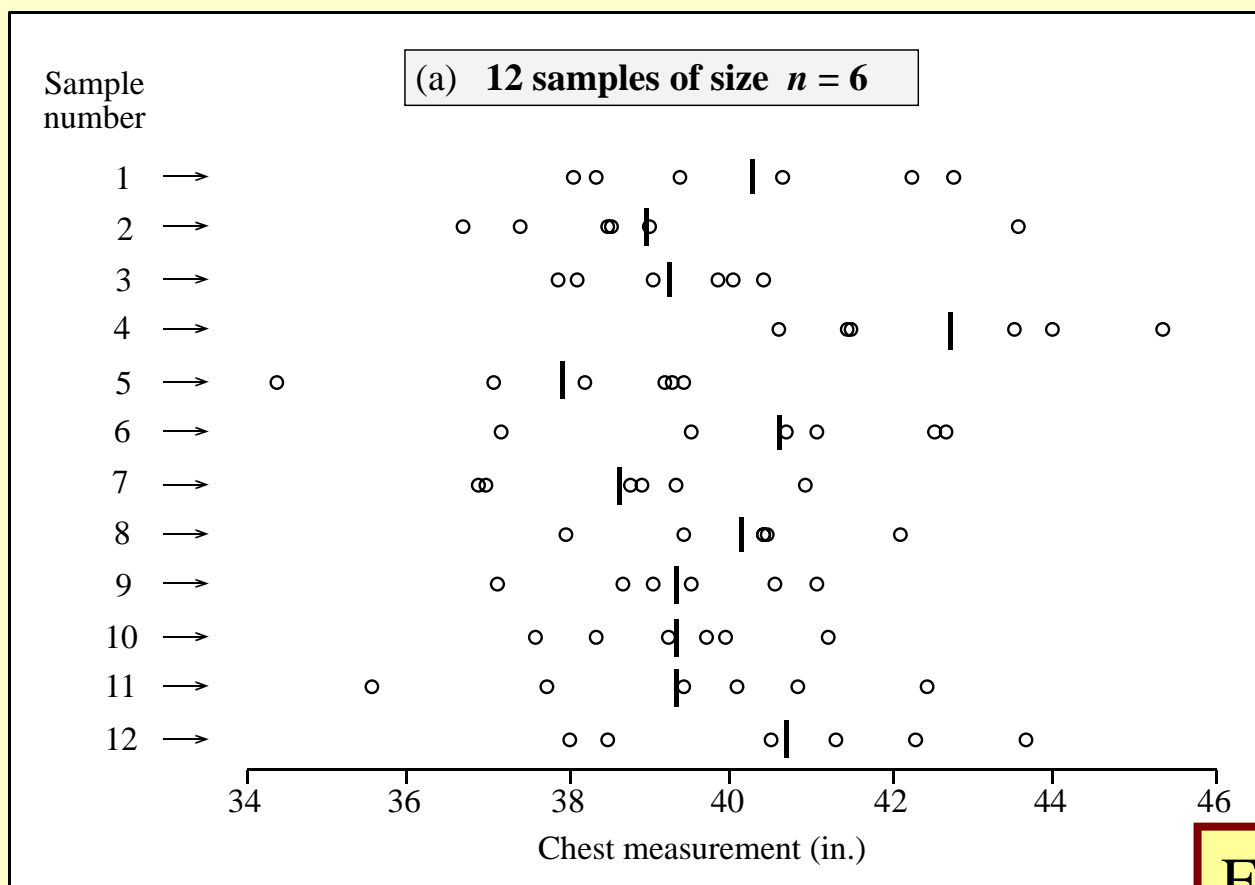


Fig. 7.2.1 (a)

Twelve samples of size 24

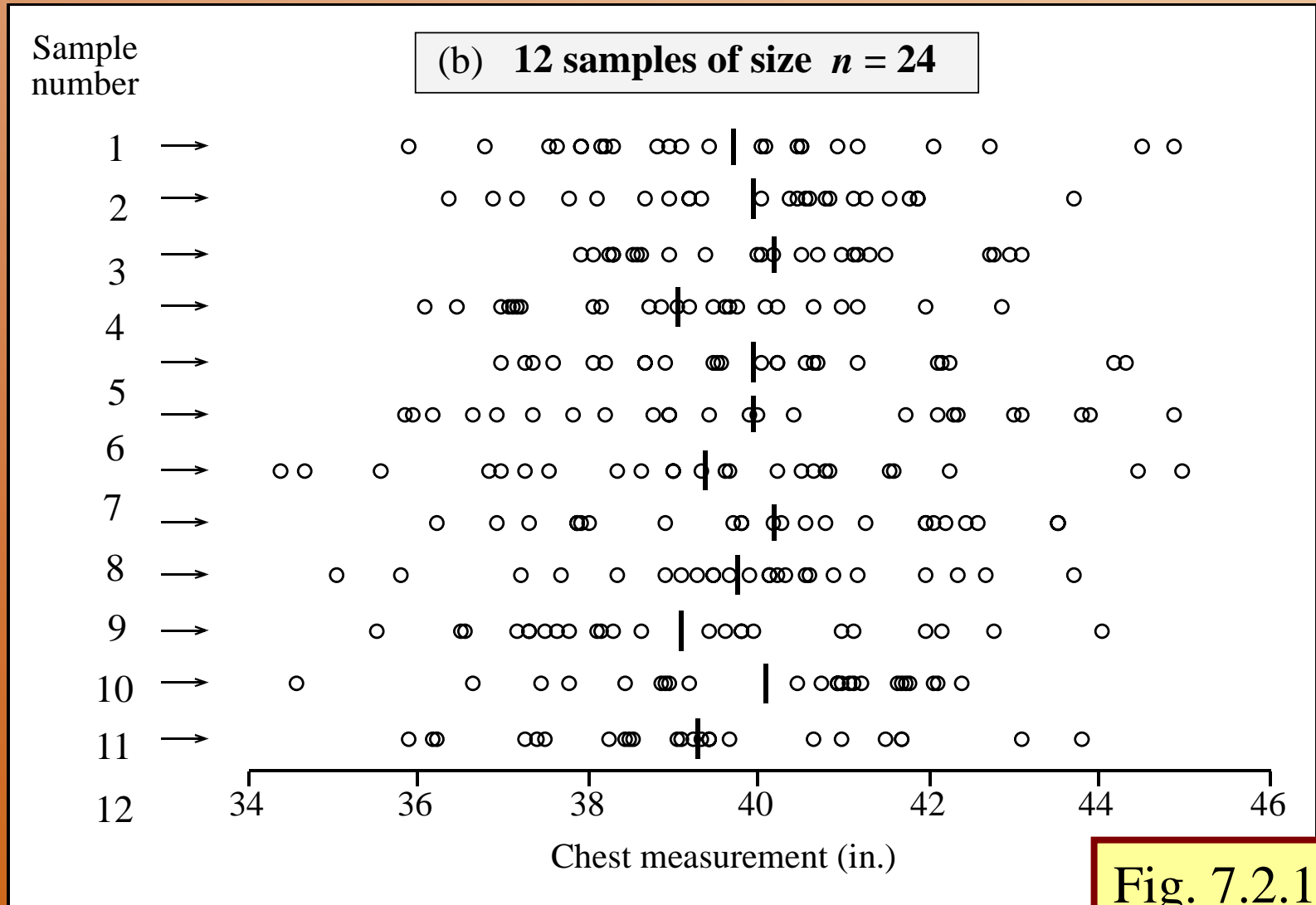


Fig. 7.2.1 (b)

Histograms from 100,000 samples

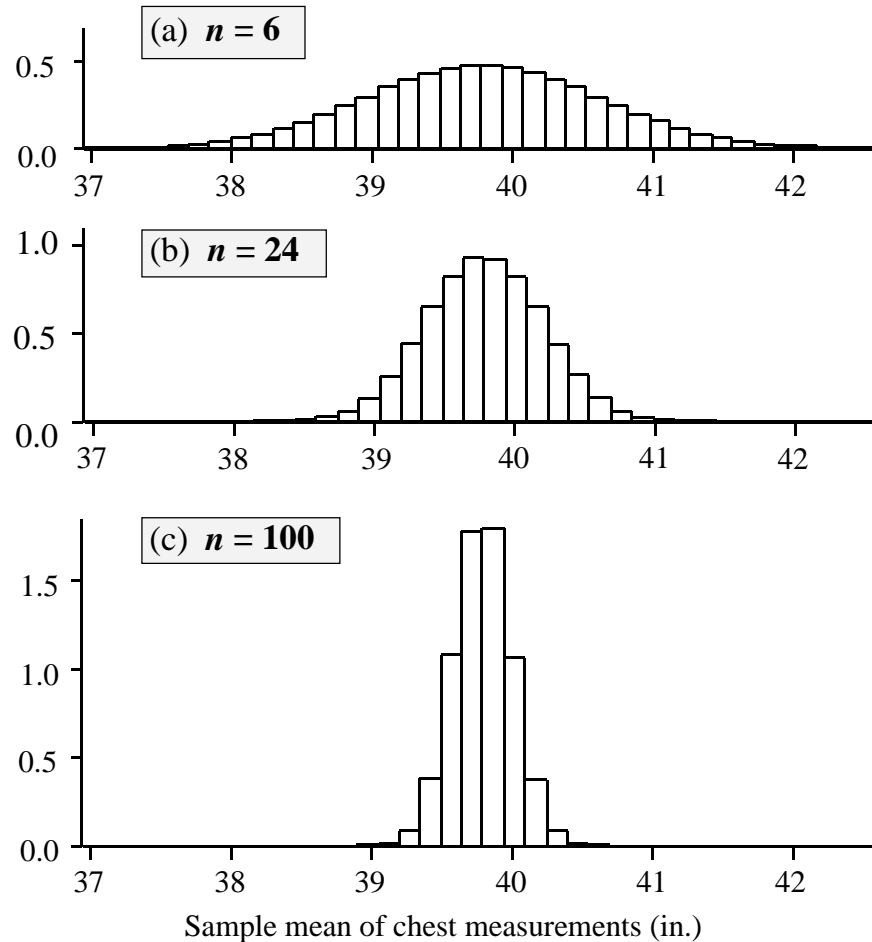
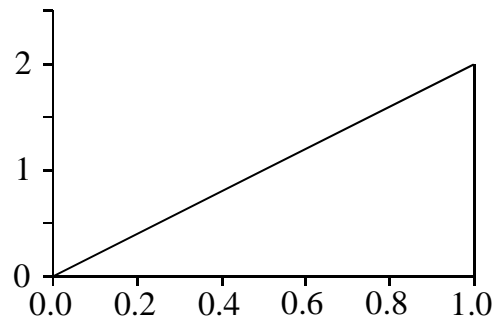


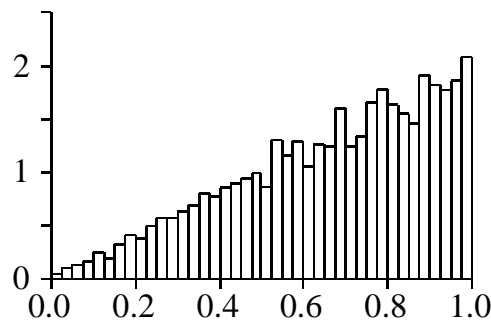
Figure 7.2.2 Standardised histograms of the sample means from 100,000 samples of soldiers (n soldiers per sample).

Central Limit Effect -- Histograms of sample means

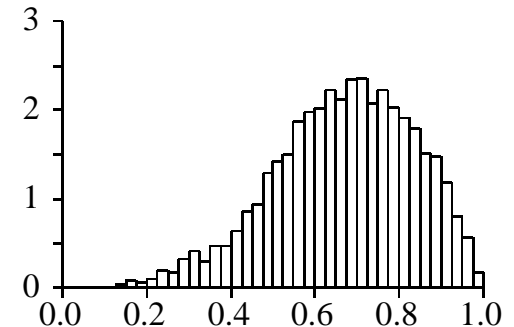
(a) **Triangular**



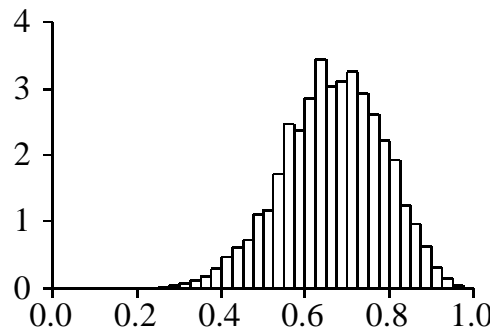
$n = 1$



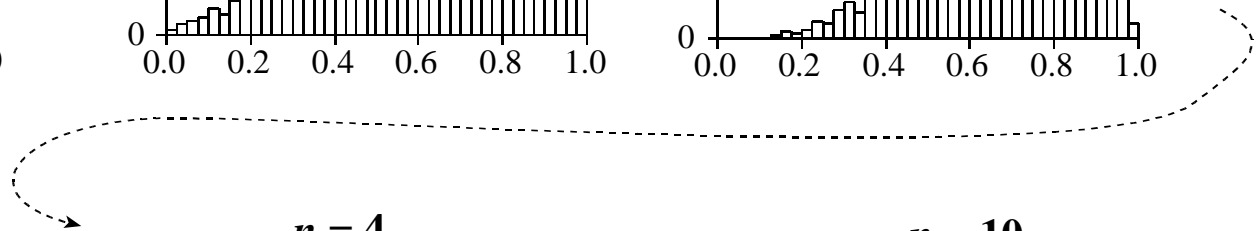
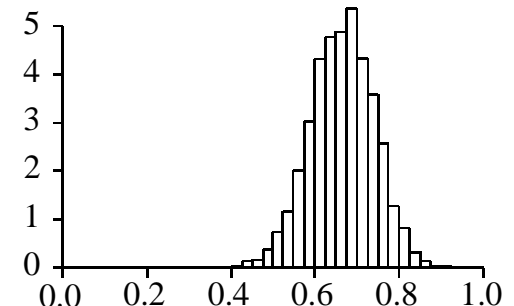
$n = 2$



$n = 4$

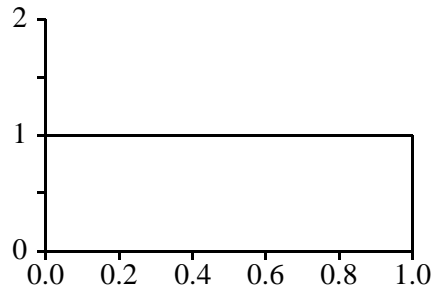


$n = 10$

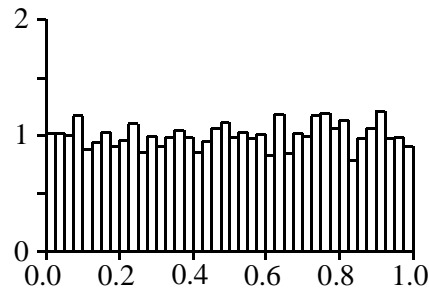


Central Limit Effect -- Histograms of sample means

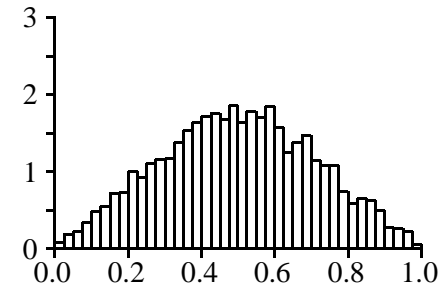
(b) Uniform



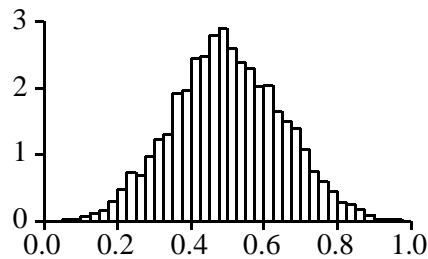
$n = 1$



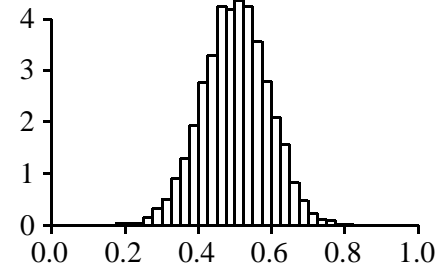
$n = 2$



$n = 4$

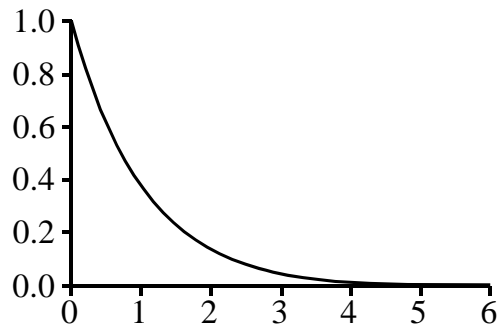


$n = 10$

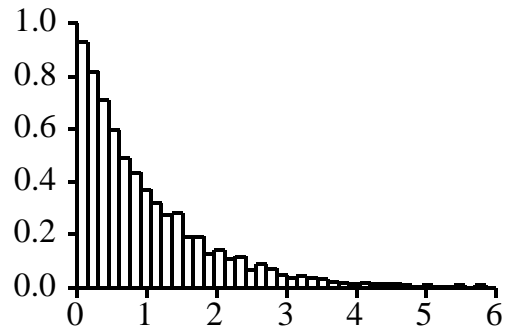


Central Limit Effect -- Histograms of sample means

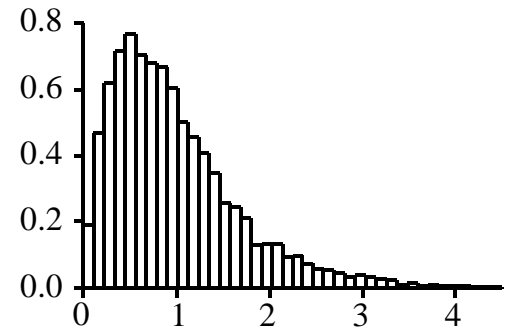
(a) Exponential



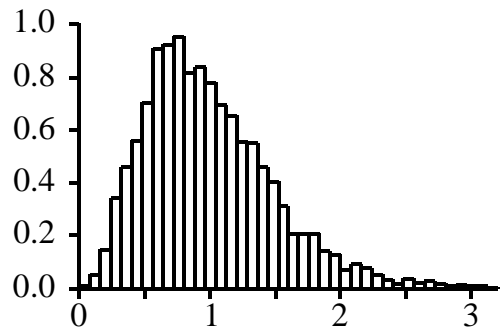
$n = 1$



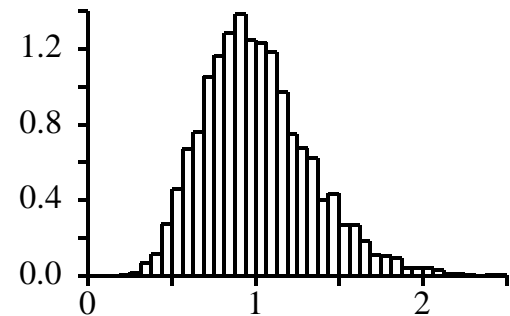
$n = 2$



$n = 4$

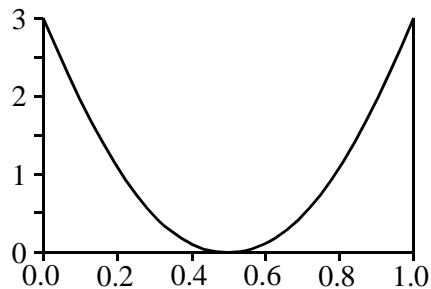


$n = 10$

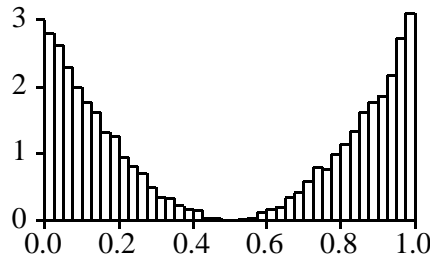


Central Limit Effect -- Histograms of sample means

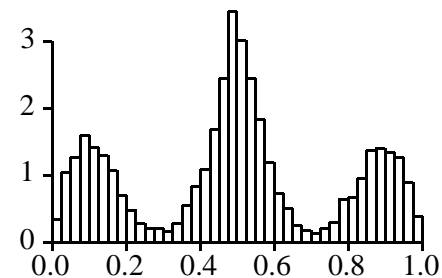
(b) Quadratic U



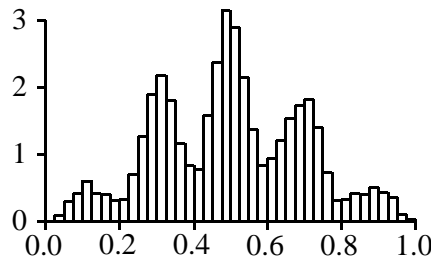
$n = 1$



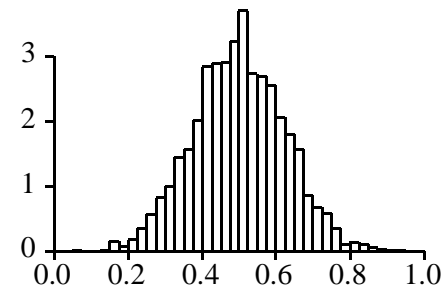
$n = 2$



$n = 4$



$n = 10$



Central Limit Theorem:

When sampling from almost any distribution,

\bar{X} is approximately Normally distributed in large samples.

Part I

Expected value

- Expected value tells us about long-run behavior in many repetitions of an experiment

$$E(X) = \sum_i X_i \cdot \text{pr}(X_i)$$

- In the continuous case

$$E(X) = \int_{-\infty}^{+\infty} X \cdot \text{pr}(X) dX$$

- Note that this is the population mean

$$\mu_X = E(X)$$

Mean and std dev of the sampling distribution

Central Limit Theorem: Part II

$$E(\text{sample mean}) = \text{Population mean}$$

$$\text{sd}(\text{sample mean}) = \frac{\text{Population standard deviation}}{\sqrt{\text{Sample size}}}$$

Standard error of the mean (SEM)

$$E(\bar{X}) = E(X) = \mu, \quad \text{sd}(\bar{X}) = \frac{\text{sd}(X)}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

Sampling distribution of the mean

- If random samples of size n are drawn from a **normal** population, the means of these samples will conform to a normal distribution
- Normal deviate of means $Z = \frac{X_i - \mu}{\sigma}$
allows us to ask probability statements about means
- What is the probability of obtaining a random sample of nine measurements with a mean larger than 50 mm from a population having a mean of 47 mm and an SD of 12 mm? (if the answer is not obvious, try it at home...)

Population and sample

$$E(\bar{X}) = E(X) = \mu, \quad \text{sd}(\bar{X}) = \frac{\text{sd}(X)}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

- Population sigma is unknown, not very useful...
- Reasonable estimate is the **standard error of the sample mean**

The standard error of the mean

The standard error of the sample mean is

- an estimate of the std dev of the sample mean
- a measure of the precision of the sample mean as an estimate of the population mean

- given by $se(\bar{x}) = \frac{\text{Sample standard deviation}}{\sqrt{\text{Sample size}}}$

$$se(\bar{x}) = \frac{s_x}{\sqrt{n}}$$

t distribution

$$Z = \frac{X_i - \mu}{\sigma}$$

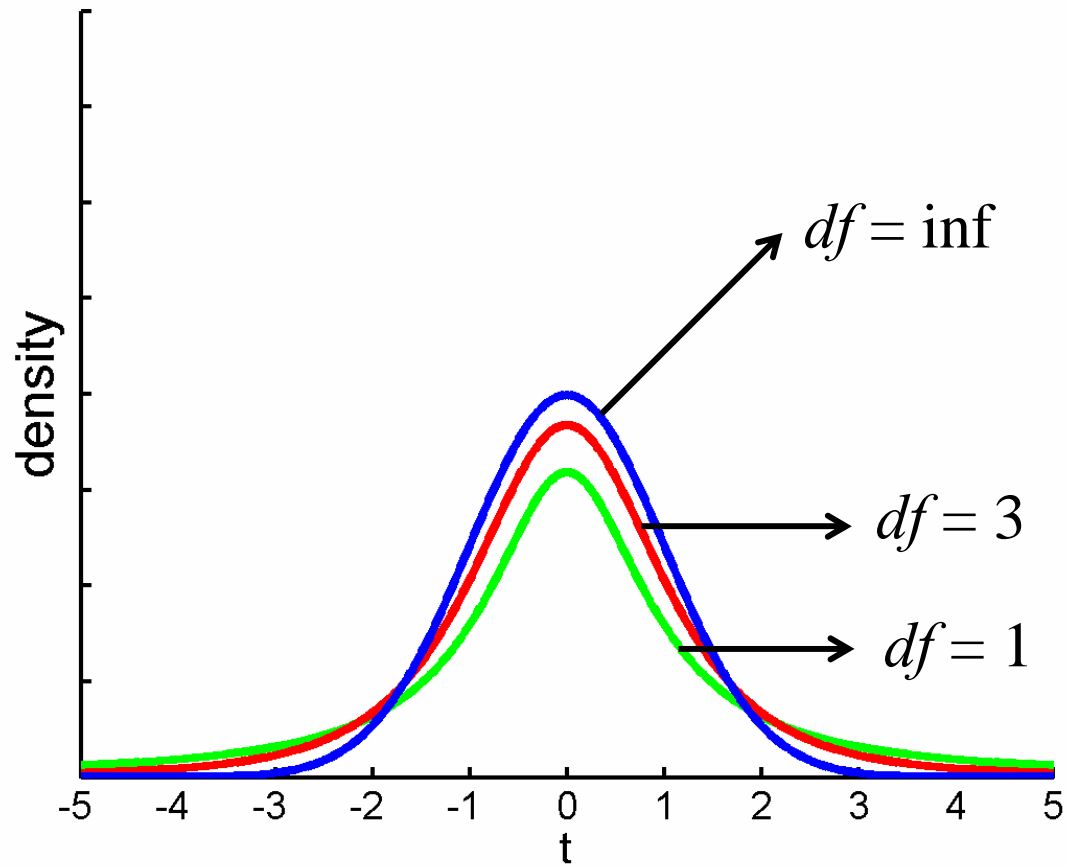
- σ is a property of the population: always unknown...
- Solution: use $s(\bar{X})$ as approximation
 - s will likely be an underestimation of σ
 - Good estimate of σ only if n is VERY large!

- Better strategy: $t = \frac{X_i - \mu}{s(\bar{X})}$

which is distributed according to a *t distribution*

- For hypotheses concerning the mean: $df = n - 1$
- $df \rightarrow \infty$ t converges to the Normal distribution

t distribution

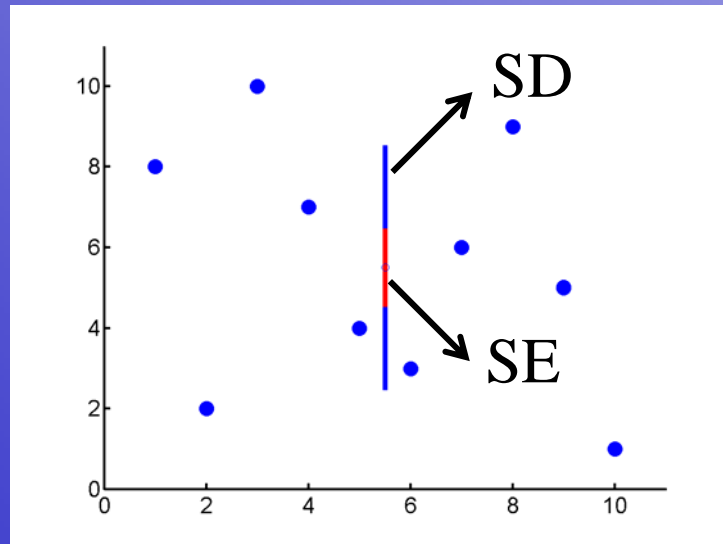


Robustness

- t distribution is obtained **only if** data are sampled from a **normal** distribution
- Fortunately, the t test is **robust**: validity is **not** seriously affected by moderate deviations from the normality assumption
 - Effect of non-normality is greater for smaller α ; effect decreases as n increases;
 - For symmetric distributions there is little effect of departure from normality;
 - Much smaller effect for two-tailed testing than for one-tailed testing.

Variability about the mean

- Reporting variability in data
- Describing the **data** being *sampled*: \bar{X} and SD
- Describing the precision of **estimation of the population mean**: SE
 - $SE < SD$



Hypothesis testing

- $H_0: \mu = 0$ (null hypothesis)
- $H_a: \mu \neq 0$ (alternative hypothesis)
- **Type I error:** probability of *rejecting* the null hypothesis when it is in fact true: α
 - Finding a difference that is not there (false positive)
- **Type II error:** probability of *not rejecting* the null hypothesis when it is in fact false: β
 - Not finding a difference when it is there (false negative)
- **Power** of a statistical test: $1 - \beta$: probability of *rejecting* the null hypothesis when it is in fact false (and should be rejected)

Type I and Type II errors

- Type I: α is the specified significance level
- Type II: β generally *unspecified* and *unknown*
- Both types of errors may be reduced *simultaneously* by increasing n

t distribution and testing the means of two independent samples

- One of the most common uses of the t test involves testing the **difference between the means of two independent groups**
- $H_0: \mu_1 - \mu_2 = 0; \mu_1 = \mu_2$
- The variance of the sum or difference of two random variables is the sum of the variances

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2$$

t distribution and testing the means of two independent samples

- From the Central Limit Theorem, we know that the variance of the distribution of \bar{X} is σ^2/N
- Thus, the standard error of the difference between means is

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

t distribution and testing the means of two independent samples

- We can write the difference in the means in terms of z scores

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

- We could compute probabilities if all values were known. But population standard deviations are rarely known...

Two independent samples

- Use *sample* standard errors as estimates of the population standard errors
- **Results will be distributed as t rather than z**

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

- Null hypothesis is generally $\mu_1 - \mu_2 = 0$

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Pooled variances

- A *very* common assumption is that $\sigma_1^2 = \sigma_2^2 = \sigma^2$
- Thus, we have two estimates of σ^2 : s_1^2 and s_2^2
- Best estimate of σ^2 is the *pooled variance* (weighted average):

$$s_p^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2} \quad N_1 + N_2 - 2 \text{ df}$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_p^2}{N_1} + \frac{s_p^2}{N_2}}} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{s_p^2 \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}}$$

Heterogeneous variances

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

- Many times the variances are not the same...**don't pool!**

$$t' = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

- But what is the distribution of t' ??
 - Homogeneity of variance is a key assumption underlying t test for two independent samples

Behrens-Fisher problem

- Determining the **sampling distribution** of t'
- Behrens and then Fischer determined distribution but only for low degrees of freedom
- Welch-Satterwaite solution: t' is distributed as t but appropriate degrees of freedom are:

$$df' = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2} \right)^2}{\frac{\left(\frac{s_1^2}{N_1} \right)^2}{N_1 - 1} + \frac{\left(\frac{s_2^2}{N_2} \right)^2}{N_2 - 1}}$$

$$\text{Min}(N_1 - 1, N_2 - 1) \leq df' \leq (N_1 + N_2 - 2)$$

Robustness

- What if we assume homogeneous variances but they are not...
- For equal samples sizes effects are quite small: within ± 0.02 of true α
- With unequal sample sizes, results are less clear cut...*use nonparametric approaches*

Parameters and estimates

- A *parameter* is a numerical characteristic of a population or distribution
- An *estimate* is a known quantity calculated from the data to estimate an unknown parameter
 - For general discussions about parameters and estimates, we talk in terms of $\hat{\theta}$ being an estimate of a parameter θ
 - The *bias* in an estimator is the difference between $E(\hat{\Theta})$ and θ
 - $\hat{\theta}$ is an *unbiased estimate* of θ if $E(\hat{\Theta}) = \theta$.

Standard error

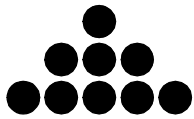
- **The standard error**, $se(\hat{\theta})$, for an estimate $\hat{\theta}$ is:
 - an estimate of the std dev. of the sampling distribution
 - a measure of the **precision** of $\hat{\theta}$ as an estimate of θ
- **For a mean**
 - The sample mean \bar{x} is an unbiased estimate of the population mean μ
 - $se(\bar{x}) = \frac{s_x}{\sqrt{n}}$

Standard error of an estimate

The *standard error* of any estimate $\hat{\theta}$ denoted $se(\hat{\theta})$

- estimates the variability of $\hat{\theta}$ values in repeated sampling and
- is a measure of the *precision* of $\hat{\theta}$

CHANCE ENCOUNTERS



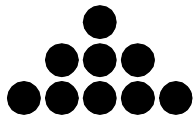
value of parameter

(a) No bias, high precision



value of parameter

(b) No bias, low precision



value of parameter

(c) Biased, high precision



value of parameter

(d) Biased, low precision

Figure 7.4.1 Bias and precision.

CHANCE ENCOUNTERS

TABLE 7.7.1 Some Parameters and Their Estimates

	Population(s) or Distributions(s) ↓ Parameters	Sample data ↓ Estimates	→ Measure of precision
Mean	μ	\bar{x}	se (\bar{x})
Proportion	p	\hat{p}	se (\hat{p})
Difference in means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	se ($\bar{x}_1 - \bar{x}_2$)
Difference in proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	se ($\hat{p}_1 - \hat{p}_2$)
General case	θ	$\hat{\theta}$	se ($\hat{\theta}$)