Sections 5.1 and 5.2; review for Exam I

Note made by: Timothy Hanson
Instructor: Peijie Hou

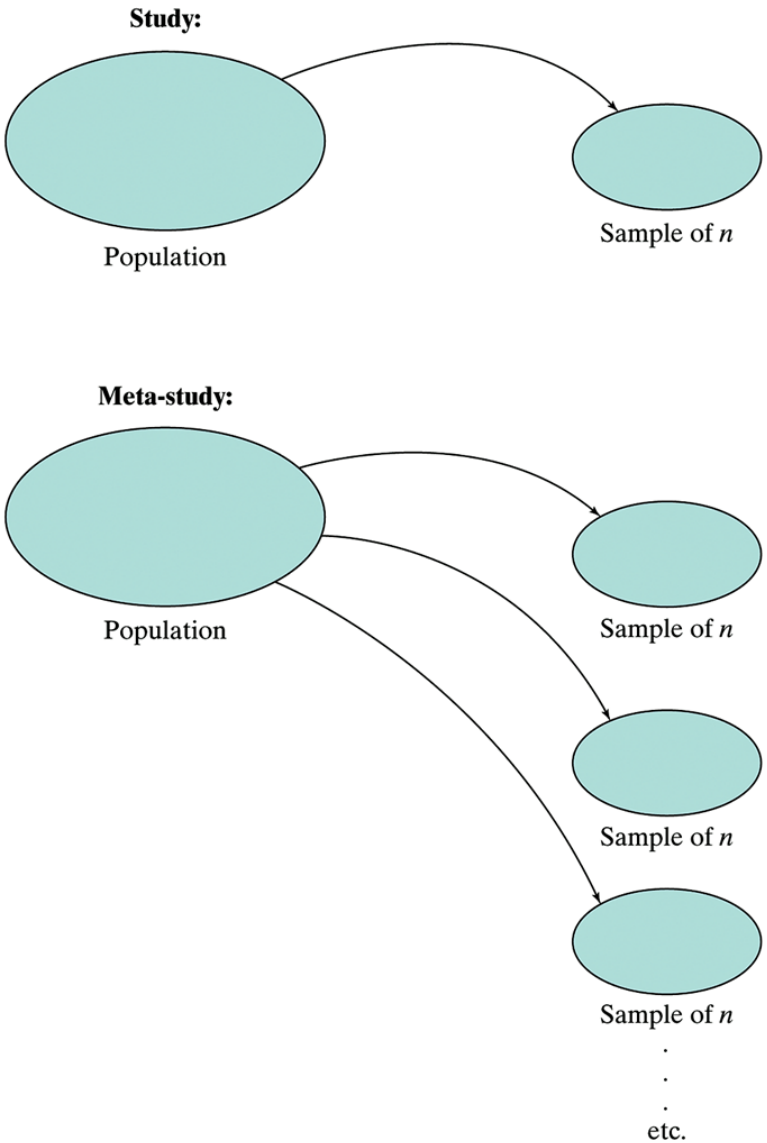
Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

Sampling variability

- A random sample is exactly that: *random*.
- You can collect a sample of n observations and compute the mean $\bar{Y}$. Before you do it, $\bar{Y}$ is random.
- If you you randomly sample a population two different times, taking, e.g. $n = 5$ each time, the two sample means $\bar{Y}_1$ and $\bar{Y}_2$ will be different.
- Example: sampling $n = 5$ ages from Stat 205.
- Variability among random samples is called **sampling variability**.
- Variability is assessed through a hypothetical "mind experiment" called a **meta-study**.

Study and meta-study



Example 5.1.1 Rat blood pressure

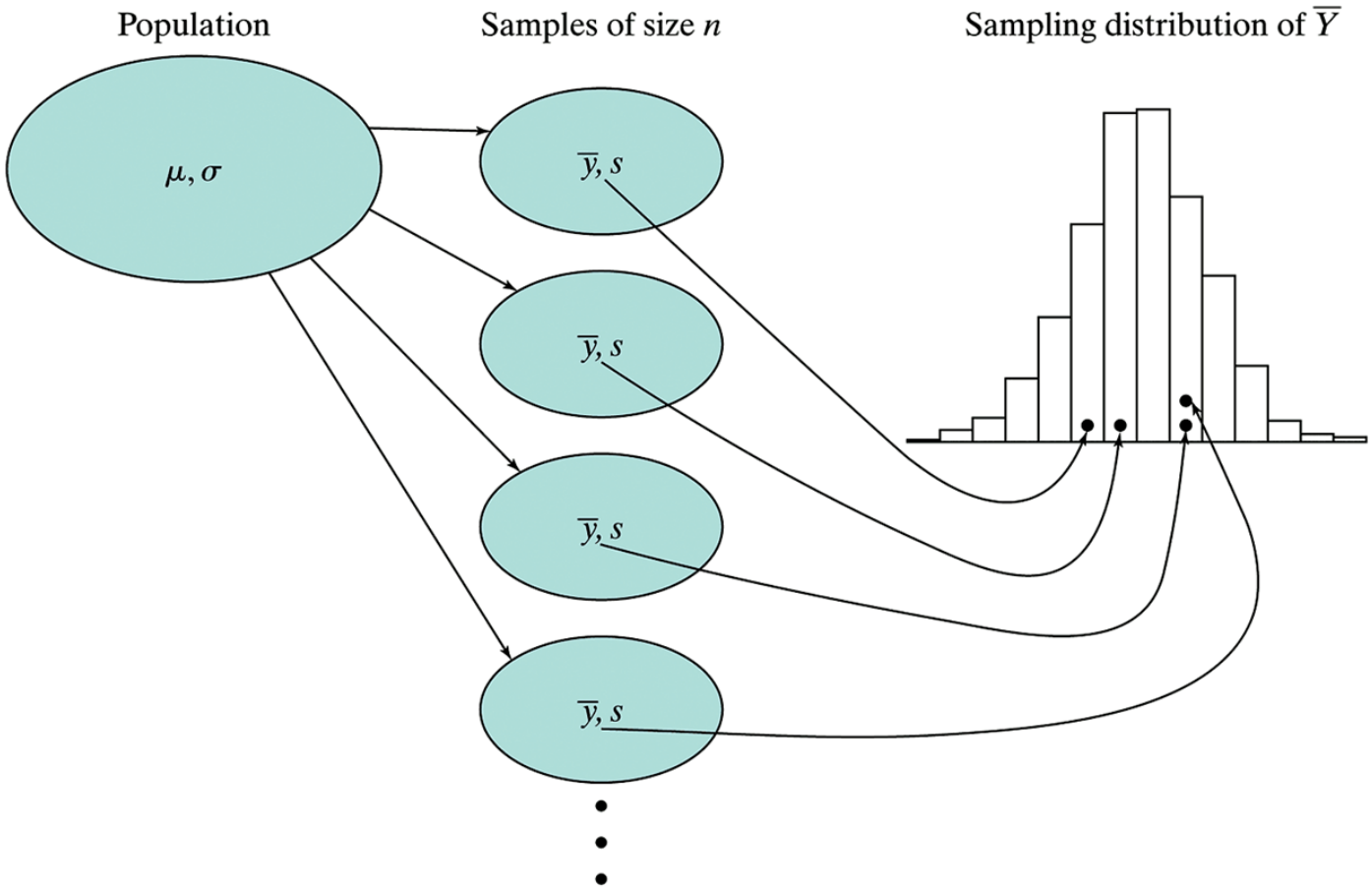
- Study is measuring change in blood pressure in $n = 10$ rats after giving them a drug, and computing a mean change $\bar{Y}$ from $Y_1, \ldots, Y_{10}$.
- Meta study (which takes place in our mind) is simply repeating this study over and over again on different samples of $n = 10$ rats and computing a mean each time $\bar{Y}_1, \bar{Y}_2, \bar{Y}_3, \ldots$
- Because the sample is random each time, the means will be different.
- A (hypothetical) histogram of the $\bar{Y}_1, \bar{Y}_2, \bar{Y}_3, \ldots$ would give the **sampling distribution** of $\bar{Y}$, and smoothed version would give the density of $\bar{Y}$.
- Restated: the sample mean *from one randomly drawn sample of size* $n = 10$ has a density.

The density of $\bar{Y}$

- $\bar{Y}$ estimates $\mu_Y = E(Y_i)$, the mean of all the observations in the population.
- We'll first look at a picture of where the **sampling distribution of** $\bar{Y}$ comes from.
- Then we'll discuss a Theorem that tells us about the mean $\mu_{\bar{Y}}$, standard deviation $\sigma_{\bar{Y}}$, and shape of the density for $\bar{Y}$.

Sampling distribution of $\bar{Y}$

"Meta-experiment..."



Sampling distribution of $\bar{Y}$

Theorem 5.2.1: The Sampling Distribution of $\overline{Y}$

1. Mean The mean of the sampling distribution of $\overline{Y}$ is equal to the population mean. In symbols,

$$\mu_{\overline{Y}} = \mu$$

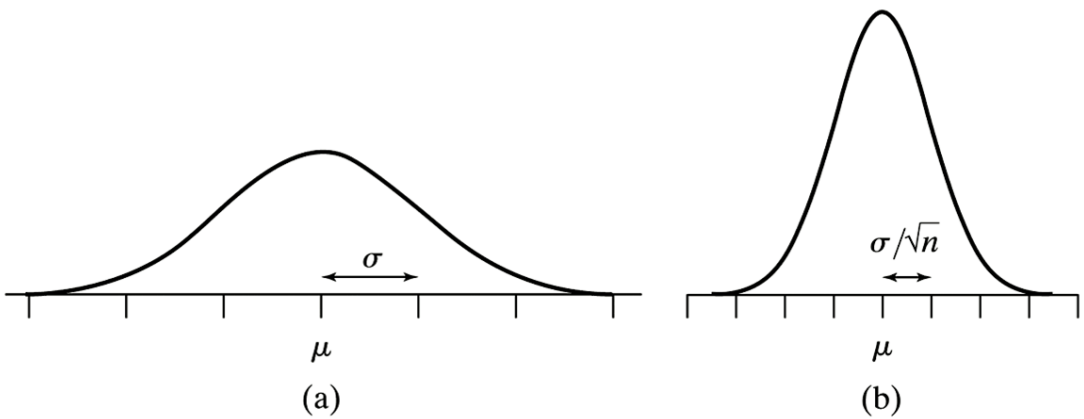
2. Standard deviation The standard deviation of the sampling distribution of $\overline{Y}$ is equal to the population standard deviation divided by the square root of the sample size. In symbols,

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$$

3. Shape
   (a) If the population distribution of Y is normal, then the sampling distribution of $\overline{Y}$ is normal, regardless of the sample size n.
   (b) *Central Limit Theorem* If n is large, then the sampling distribution of $\overline{Y}$ is approximately normal, even if the population distribution of Y is not normal.

Sampling distribution of $\bar{Y}$ from normal data

If data $Y_1, Y_2, \ldots, Y_n$ are normal, then $\bar{Y}$ is *also normal*, centered at the same place as the data, but with smaller spread.



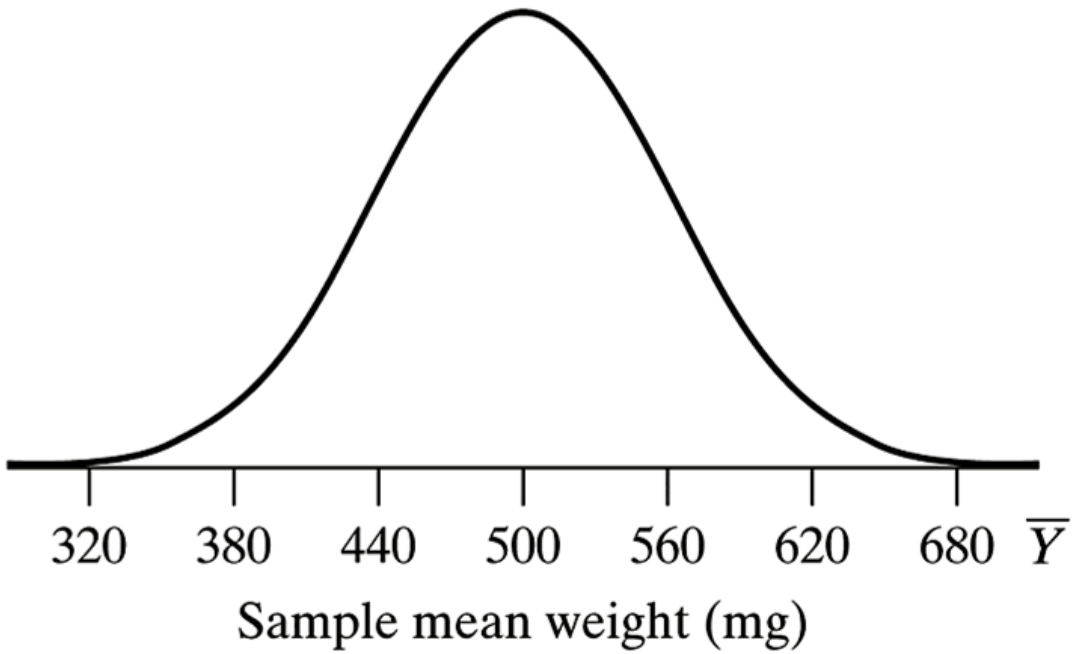
(a) population distribution of normal data $Y_1, \ldots, Y_n$, and (b) sampling distribution of $\bar{Y}$.

Example 5.2.2 Seed weights

- The population of weights of the princess bean is *normal* with $\mu = 500$ mg and $\sigma = 120$ mg. We intend to take a samplle of $n = 4$ seeds and compute the (random!) sample mean $\bar{Y}$.
- $E(\bar{Y}) = \mu_{\bar{Y}} = \mu = 500$ mg. *On average, the sample mean gets it right*.
- $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{120}{\sqrt{4}} = 60$ mg. 68% of the time, $\bar{Y}$ will be within 60 mg of $\mu = 500$ mg.

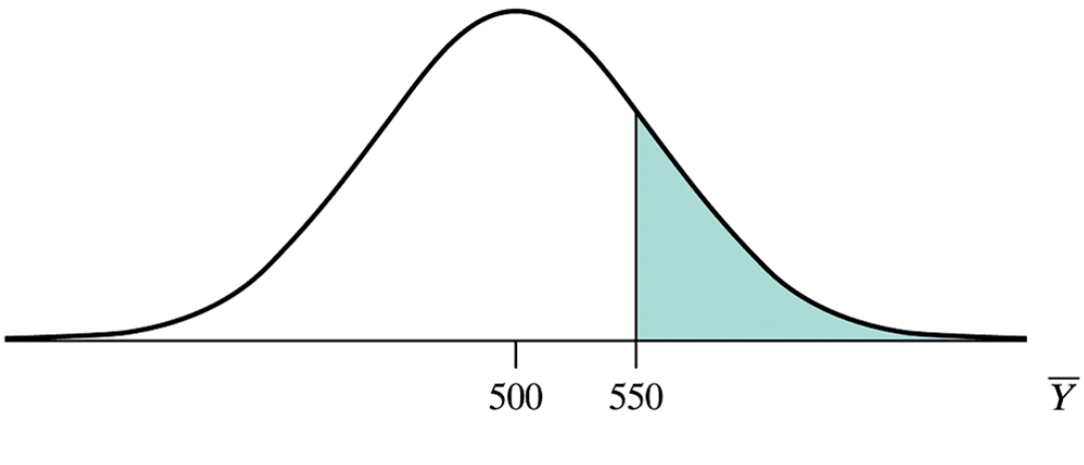
Sampling distribution for $\bar{Y}$ for Example 5.2.2

$\mu_{\bar{Y}} = 500$ mg and $\sigma_{\bar{Y}} = 60$ mg.



$\Pr\{\bar{Y} > 550\}$ for $n = 4$

Recall for $n = 4$ that $\mu_{\bar{Y}} = 500$ mg and $\sigma_{\bar{Y}} = 60$ mg.

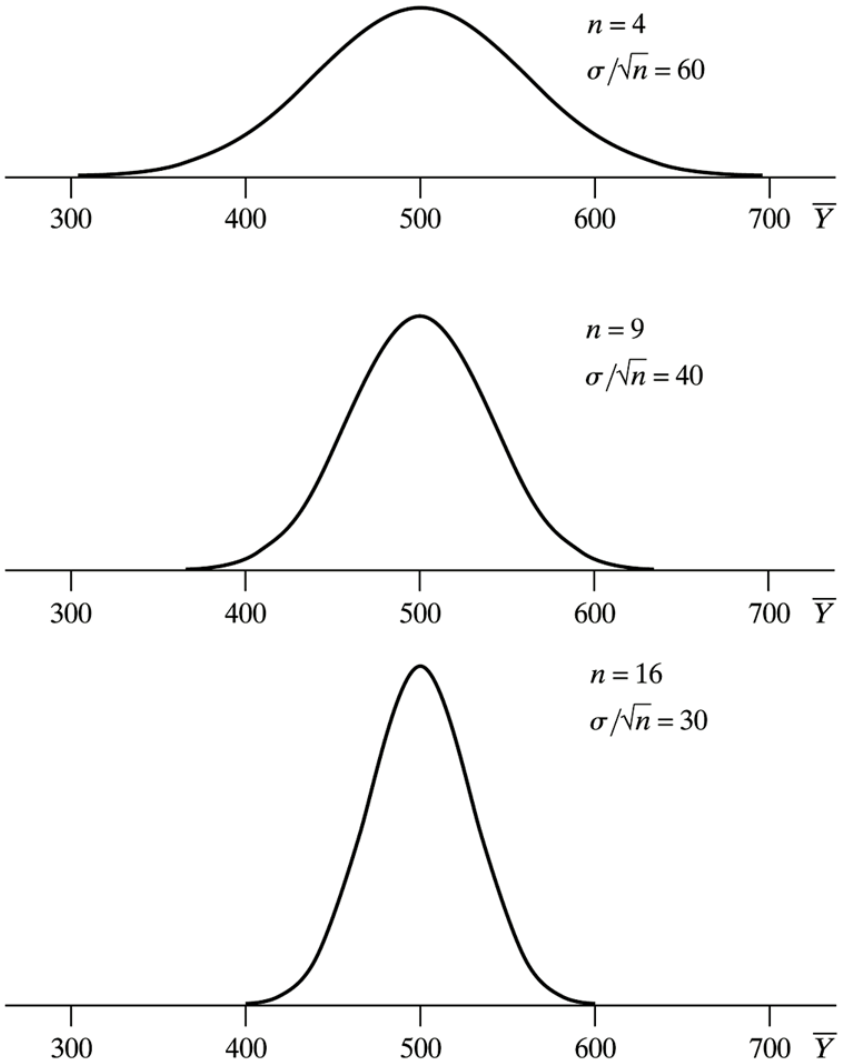



```
> 1-pnorm(550,500,60)
[1] 0.2023284
```

What happens when n is increased?

- As n gets bigger, $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$ gets smaller. The density of $\bar{Y}$ gets more focused around μ .
- If $Y_1, \ldots, Y_n$ come from a normal density, then so does $\bar{Y}$, *regardless of the sample size*.
- Even if $Y_1, \ldots, Y_n$ *do not* come from a normal density, the *Central Limit Theorem* guarantees that the density of $\bar{Y}$ will look more and more like a normal distribution as n gets bigger.
- This is in Section 5.3; have a look if you're interested.

Sampling dist'n for $\bar{Y}$ from different sample sizes n



Exam I logistics...

- Tuesday, February 17, 8:30–9:45.
- Closed book, closed notes. Bring a calculator and a pencil.
- Problems will be patterned after homework problems; multiple choice.
- *Be on time*.
- Note that textbook problem solutions manual are posted on the course website.
- Each of exams I, II, and III are worth 15% of your final grade.
- No phones, no hats. The exam is straightforward; I will not try to "trick" you.

2.1 Types of variables

- Categorical
  - Ordinal (e.g. "low, medium, high", "infant, toddler, child, teen, adult")
  - Nominal (e.g. eye color, car type)
- Numeric
  - Continuous (e.g. height, cholesterol, tree diameter)
  - Discrete (e.g. number of cracked eggs in a carton, die roll)

HW 2.1.1, 2.1.2, 2.1.4

2.2: Histograms, distributions, skew and modality

- Have data $y_1, y_2, \ldots, y_n$; want to describe it with pictures and tables.
- If data categorical, can make a bar chart. Can record frequency of data value occurrences in a table.
- Continuous data can be displayed in a histogram defined by bins. Again, need a table of frequency values for occurrences within each bin.
- Histogram/density shape: unimodal, bimodal, multimodal.
- Histogram/density skew: left skew, right skew, symmetry.
- HW 2.2.3 (use R).

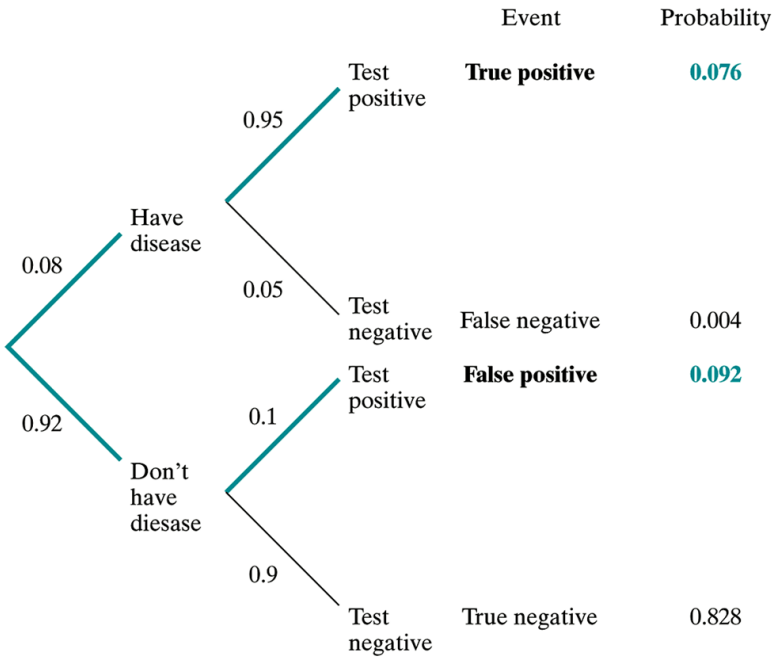
2.3, 2.4, 2.6: Descriptive statistics: mean, median, quartiles, 5 number summary, IQR, boxplots, outliers.

- Mean $\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n)$ is "balance point" of data.
- Median Q_2 (or $\tilde{y}$) cuts ordered data into halves of equal size.
- First quartile Q_1 is median of lower half; Third quartile Q_3 is median of upper half.
- min, $Q_1, \tilde{y}, Q_3$, max is 5 number summary, used to make boxplot.
- $IQR = Q_3 - Q_1$, length of interval containing middle 50% of data. Sample variance is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, standard deviation is s.
- $UF = Q_3 + 1.5 \times IQR$, $LF = Q_1 - 1.5 \times IQR$. Any of $y_1, \ldots, y_n$ **larger than** UF or **smaller than** LF are "outliers."
- HW 2.3.4 (use R), 2.4.2 (use R), 2.4.7, 2.6.5 (by hand), 2.6.11, 2.6.12.

3.3: Probability

- Let A and B two events. A and B is that both occur. A or B is either occurs. A^C is that A does not occur. *Always*: $0 \leq \Pr\{A\} \leq 1$.
- A and B are *disjoint* if they have no outcomes in common.
- Formulas:
  1. If $E_1, E_2, \ldots, E_k$ disjoint, then $\Pr\{E_1 \text{ or } E_2 \text{ or } \cdots \text{ or } E_k\} = \Pr\{E_1\} + \Pr\{E_2\} + \cdots + \Pr\{E_k\}$.
  2. $\Pr\{A \text{ or } B\} = \Pr\{A\} + \Pr\{B\} - \Pr\{A \text{ and } B\}$.
  3. (conditional probability) $\Pr\{A|B\} = \Pr\{A \text{ and } B\}/\Pr\{B\}$.
  4. (compliment rules) $\Pr\{A^C\} = 1 - \Pr\{A\}$ *and* $\Pr\{A^C|B\} = 1 - \Pr\{A|B\}$.
  5. (independence) A and B are independent if $\Pr\{A\} = \Pr\{A|B\}$.
- HW 3.3.1, 3.3.2, 3.3.3, 3.3.4.

3.2 Probability trees



$\Pr\{\text{Disease, Test positive}\} = 0.08(0.95) = 0.076$ $\Pr\{\text{Disease, Test negative}\} = 0.08(0.05) = 0.004$

$\Pr\{\text{No disease, Test positive}\} = 0.92(0.10) = 0.092$ $\Pr\{\text{No disease, Test negative}\} = 0.92(0.90) = 0.828$

What is the probability of testing positive? What is $\Pr\{\text{disease}|\text{test positive}\}$? HW 3.2.3, 3.2.4, 3.2.5, 3.2.7

3.4: Continuous random variables, densities

- A continuous random variable Y has a *density* $f(y)$. Examples: cholesterol, height, GPA, blood pressure.
- $\Pr\{a < Y < b\}$ is the area under the density curve $f(y)$ between a and b. Total area equals one.
- Note that $\Pr\{Y < a\} = \Pr\{Y \leq a\}$. Only with *continuous* random variables.
- HW 3.4.2, 3.4.3.

3.5: Discrete random variables

- A *discrete* random variable can only take on a countable number of values. Examples: number of broken eggs in a carton, number of earthquakes in a day.
- Finite discrete random variables have probability mass functions, e.g.

No. vertebrae y	$\Pr\{Y = y\}$
20	0.03
21	0.51
22	0.40
23	0.06

- Get probabilities $\Pr\{a \leq Y \leq b\}$ by summing probabilities in table for $a \leq y \leq b$.
- For discrete $\Pr\{Y < a\}$ will be different than $\Pr\{Y \leq a\}$.

3.5: Discrete random variables

- Mean is now *weighted average*

$$\mu_Y = E(Y) = \sum y_i \Pr\{Y = y_i\}.$$

- Variance is *weighted average* squared deviation about mean

$$\sigma_Y^2 = \sum (y_i - \mu_Y)^2 \Pr\{Y = y_i\}.$$

- Standard deviation is σ_Y.
- HW 3.5.4, 3.5.5.

3.6: Binomial distribution

- Notation $Y \sim \text{binomial}(n, p)$. Y counts number of "success" trials out of n. Y can be $0, 1, 2, \ldots, n$.
- $\Pr\{Y = j\} = {}_nC_j \; p^j (1-p)^{n-j}$ for $j = 0, 1, \ldots, n$.
- I will give you R output for $\Pr\{Y = 0\}, \Pr\{Y = 1\}, \ldots, \Pr\{Y = n\}$
- $\mu_Y = E(Y) = n\ p$, $\sigma_Y^2 = n\ p\ (1-p)$.
- HW 3.6.1, 3.6.2, 3.6.6, 3.6.10 (use R for all of these).

4.2, 4.3: Normal distribution

- Used to model *many, many* different kinds of continuous data: cholesterol, eggshell thickness, creatinine clearance, $T_{1\rho}$ measurements from MRI, health care expenditures, etc.
- Notation: $Y \sim N(\mu, \sigma^2)$.
- μ is mean and σ^2 is variance of Y (requires calculus to show this). σ is standard deviation.
- Y is *continuous* random variable that can be any number $-\infty < Y < \infty$.
- Get probabilities from R using pnorm(y,μ,σ).

Normal distribution

μ and σ *are given to you* in Chapter 4.

$$\begin{aligned} \Pr\{a < Y < b\} &= \text{pnorm}(b,\mu,\sigma) - \text{pnorm}(a,\mu,\sigma) \\ \Pr\{Y < b\} &= \text{pnorm}(b,\mu,\sigma) \\ \Pr\{Y > a\} &= 1 - \text{pnorm}(a,\mu,\sigma) \end{aligned}$$

Normal distribution

- Let $Y \sim N(\mu, \sigma^2)$. Say we want y^* such that $\Pr\{Y < y^*\} = p$ where p is given.
- qnorm(p,μ,σ) gives y^*.
- y^* is called $p(100)$th percentile of Y.
- e.g. If $Pr\{Y \leq 10\} = 0.7$ then the $70th$ percentile of Y is 10.
- HW 4.3.3, 4.4.4, 4.3.5, 4.3.6, 4.3.8 (use R for all of these).