Power and Sample Size Determination

Bret Hanlon and Bret Larget

Department of Statistics
University of Wisconsin—Madison

November 3–8, 2011

Experimental Design

- To this point in the semester, we have largely focused on methods to analyze *the data that we have* with little regard to *the decisions on how to gather the data.*

- *Design of Experiments* is the area of statistics that examines plans on how to gather data to achieve good (or optimal) inference.

- Here, we will focus on the question of sample size:
  - ▶ how large does a sample need to be so that a confidence interval will be no wider than a given size?
  - ▶ how large does a sample need to be so that a hypothesis test will have a low p-value if a certain alternative hypothesis is true?

- Sample size determination is just one aspect of good design of experiments: *we will encounter additional aspects in future lectures.*

Proportions

- Recall methods for inference about proportions: confidence intervals

Confidence Interval for p

A P% confidence interval for p is

$$p' - z^* \sqrt{\frac{p'(1 - p')}{n'}} < p < p' + z^* \sqrt{\frac{p'(1 - p')}{n'}}$$

where $n' = n + 4$ and $p' = \frac{X+2}{n+4} = \frac{X+2}{n'}$ and z^* is the critical number from a standard normal distribution where the area between $-z^*$ and z^* is $P/100$. (For 95%, $z^* = 1.96$.)

Proportions

- ...and hypothesis tests.

The Binomial Test

If $X \sim \mathrm{Binomial}(n, p)$ with null hypothesis $p = p_0$ and we observe $X = x$, the p-value is the probability that a new random variable $Y \sim \mathrm{Binomial}(n, p_0)$ would be at least as extreme (either $P(Y \leq x)$ or $P(Y \geq x)$ or $P(|Y - np_0| \geq |x - np_0|)$ depending on the alternative hypothesis chosen.)

Sample size for proportions

Case Study

- Next year it is likely that there will be a recall election for Governor Scott Walker.

- A news organization plans to take a poll of likely voters over the next several days to find, if the election were held today, the proportion of voters who would vote for Walker against an unnamed Democratic opponent.

- Assuming that the news organization can take a random sample of likely voters:

    How large of a sample is needed for a 95% confidence interval to have a margin of error of no more than 4%?

Calculation

Example

- Notice that the margin of error depends on both n and p', but we do not know p'.

$$1.96\sqrt{\frac{p'(1 - p')}{n + 4}}$$

- However, the expression $p'(1 - p')$ is maximized at 0.5; if the value of p' from the sample turns out to be different, the margin of error will just be a bit smaller, which is even better.

- So, it is conservative (in a statistical, not political sense) to set $p' = 0.5$ and then solve this inequality for n.

$$1.96\sqrt{\frac{(0.5)(0.5)}{n + 4}} < 0.04$$

- Show on the board why $n > \left(\frac{(1.96)(0.5)}{0.04}\right)^2 - 4 \doteq 621$.

General Formula

Sample size for proportions

$$n > \left(\frac{(1.96)(0.5)}{M} \right)^2 - 4$$

where M is the desired margin of error.

Errors in Hypothesis Tests

- When a hypothesis test is used to make a decision to reject the null hypothesis when the p-value is below a prespecified fixed value α, there are two possible correct decisions and two possible errors.
- We first saw these concepts with proportions, but review them now.
- The two decisions we can make are to Reject or Not Reject the null hypothesis.
- The two states of nature are the the null hypothesis is either True or False.
- These possibilities combine in four possible ways.

	H_0 is True	H_0 is False
Reject H_0	Type I error	Correct decision
Do not Reject H_0	Correct decision	Type II error

Type I and Type II Errors

Definition
- A *Type I Error* is rejecting the null hypothesis when it is true.
- The probability of a type I error is called the *significance level of a test* and is denoted α.

Definition
- A *Type II Error* is not rejecting a null hypothesis when it is false.
- The probability of a type II error is called β, but the value of β typically depends on which particular alternative hypothesis is true.

Definition
- The *power* of a hypothesis test for a specified alternative hypothesis is $1 - \beta$.
- The power is the probability of rejecting the null hypothesis in favor of the specific alternative.

Graphs

- Note that as there are many possible alternative hypotheses, for a single α there are many values of β.

- It is helpful to plot the probability of rejecting the null hypothesis against the parameter values.

Sample size calculation

> **Example**
>
> - Consider a population with proportion p.
> - Let X be the number of successes in a random sample of size 100
> - with model $X \sim \mathrm{Binomial}(100, p)$.
> - Consider the hypotheses $H_0 : p = 0.3$ versus $H_A : p < 0.3$.
> - The researchers decide to reject the null hypothesis if $X \leq 22$.
>   1. Find α
>   2. Find β if $p = 0.2$.
>   3. Plot the probability of rejecting the null hypothesis versus p.

Calculation

Solution

$$\alpha = P(X \leq 22 \mid p = 0.3)$$

$$= \sum_{k=0}^{22} \binom{100}{k} (0.3)^k (0.7)^{100-k}$$

$$\doteq 0.0479$$

$$1 - \beta(p) = P(X \leq 22 \mid p)$$

$$= \sum_{k=0}^{22} \binom{100}{k} p^k (1-p)^{100-k}$$

$$1 - \beta(0.2) \doteq 0.7389$$

Two binomial distributions
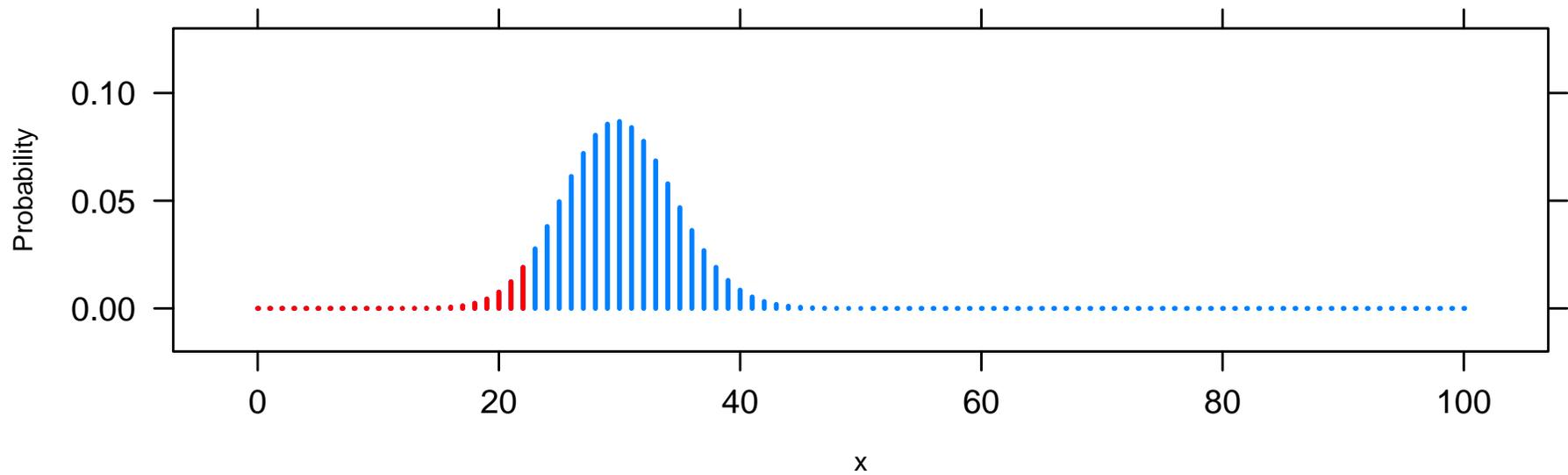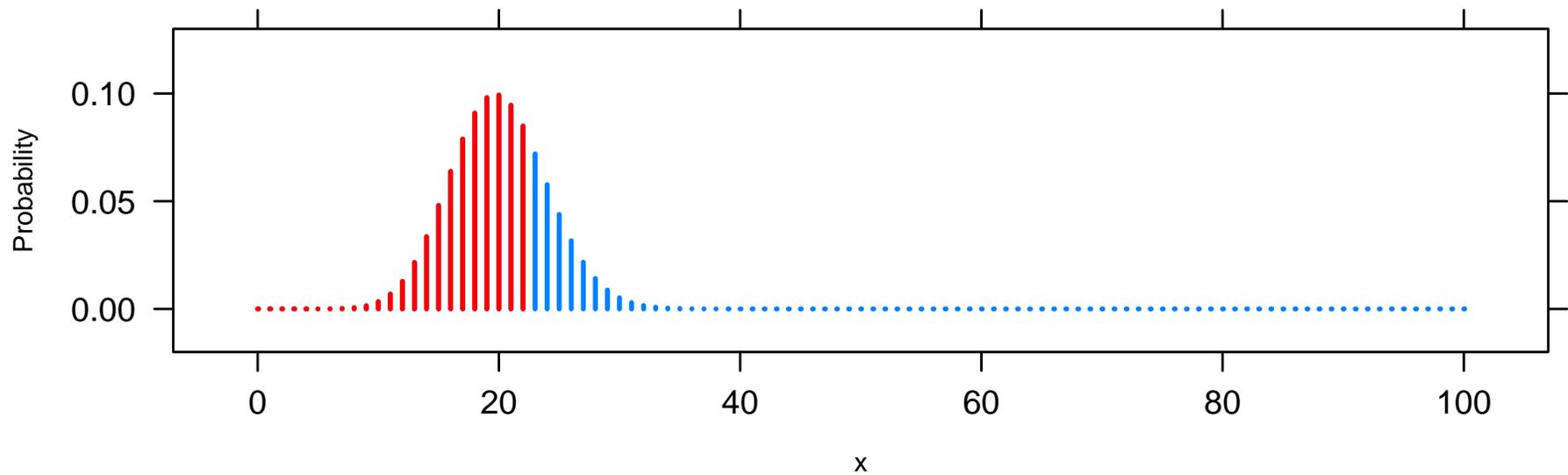
Graph of Power
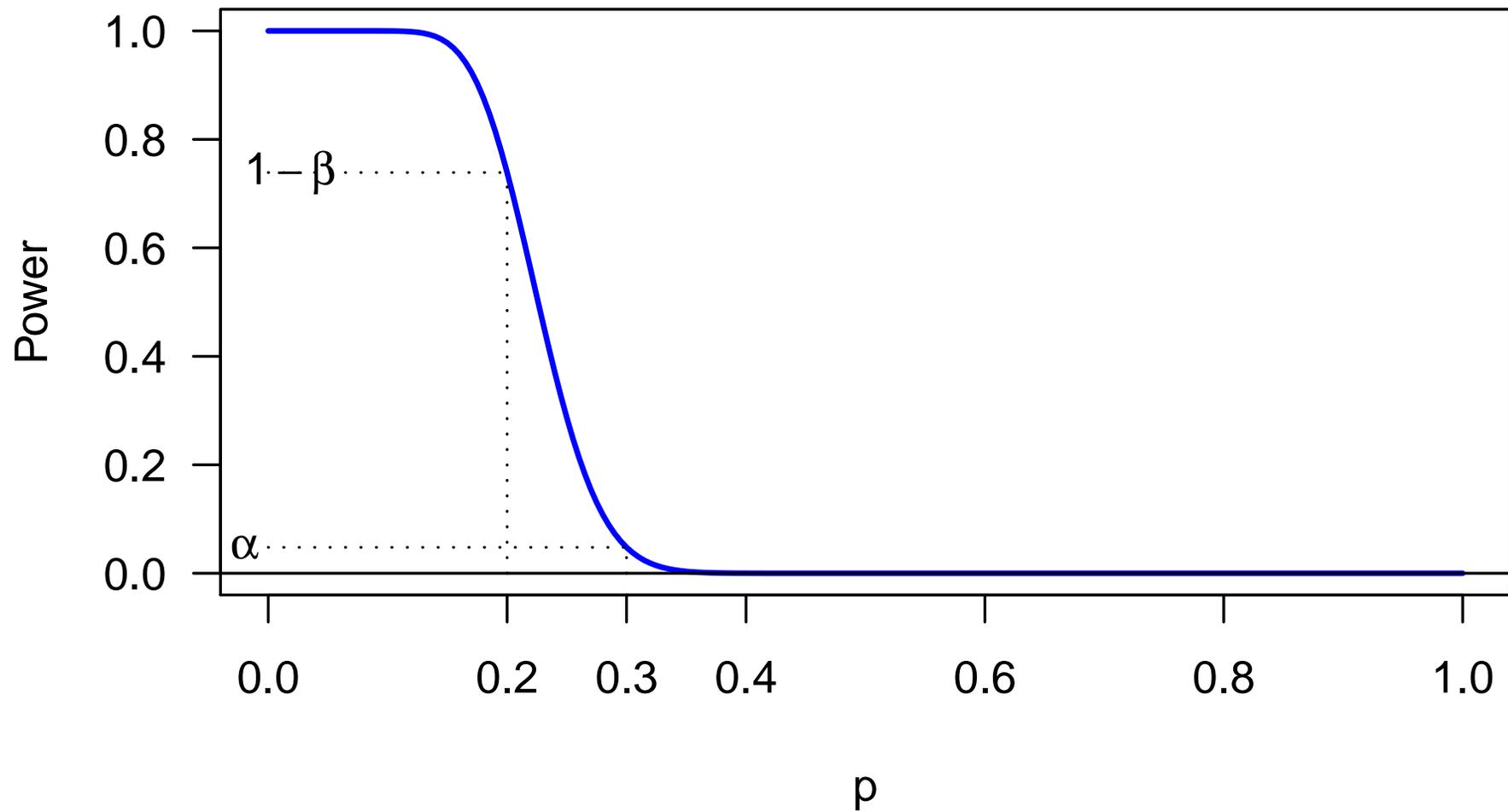
Sample Size

> **Example**
>
> - Suppose that we wanted a sample size large enough so that we could pick a rejection rule where α was less than 0.05 and the power when $p = 0.2$ was greater than 0.9.
>
> - How large would n need to be?
>
> - Simplify by letting n be a multiple of 25.
>
> - We see in the previous example that 100 is not large enough: if the critical rejection value were more than 22, then $\alpha > 0.05$ and the power is 0.7389 which is less than 0.9.
>
> - The calculation is tricky because as n changes we need to change the rejection rule so that $\alpha \leq 0.05$ and then find the corresponding power.

Calculation

- We can use the quantile function for the binomial distribution, qbinom(), to find the rejection region for a given α.
- For example, for $X \sim \text{Binomial}(100, 0.3)$,

```
> k = qbinom(0.05, 100, 0.3)
> k

[1] 23

> pbinom(k, 100, 0.3)

[1] 0.07553077

> pbinom(k - 1, 100, 0.3)

[1] 0.04786574
```

we see that
$$P(X \leq 22) < 0.05 < P(X \leq 23)$$

and rejecting the null hypothesis when $X \leq 22$ results in a test with $\alpha < 0.05$.

Calculation

- We can just substract one from the quantile function in general to find the rejection region for a given n.

  ```
  > k = qbinom(0.05, 100, 0.3) - 1
  > k
  ```

  ```
  [1] 22
  ```

- The power when $p = 0.2$ is

  ```
  > pbinom(k, 100, 0.2)
  ```

  ```
  [1] 0.7389328
  ```

Calculation

- This R code will find the rejection region, significance level, and power for $n = 100, 125, 150, 175, 200$.

```
> n = seq(100, 200, 25)
> k = qbinom(0.05, n, 0.3) - 1
> alpha = pbinom(k, n, 0.3)
> power = pbinom(k, n, 0.2)
> data.frame(n, k, alpha, power)

    n  k      alpha      power
1 100 22 0.04786574 0.7389328
2 125 28 0.03682297 0.7856383
3 150 35 0.04286089 0.8683183
4 175 42 0.04733449 0.9193571
5 200 48 0.03594782 0.9309691
```

- We see that $n = 175$ is the smallest n which is a multiple of 25 for which a test with $\alpha < 0.05$ has power greater than 0.9 when $p = 0.2$.

Normal Populations

- The previous problems were for the binomial distribution and proportions, which is tricky because of the discreteness and necessary sums of binomial probability calculations.

- Answering similar problems for normal populations is easier.

- However, we need to provide a guess for σ.

Butterfat

> **Example**
>
> - We want to know the mean percentage of butterfat in milk produced by area farms.
>
> - We can sample multiple loads of milk.
>
> - Previous records indicate that the standard deviation among loads is 0.15 percent.
>
> - How many loads should we sample if we desire the margin of error of a 99% confidence interval for μ, the mean butterfat percentage, to be no more than 0.03?

Confidence Intervals

- We will use the t distribution, not the standard normal for the actual confidence interval, but it is easiest to use the normal distribution for planning the design.

- If the necessary sample size is even moderately large, the differences between z^* and t^* is tiny.

- Recall the confidence interval for μ.

Confidence Interval for μ

A P% confidence interval for μ has the form

$$\bar{Y} - t^* \frac{s}{\sqrt{n}} < \mu < \bar{Y} + t^* \frac{s}{\sqrt{n}}$$

where t^* is the critical value such that the area between $-t^*$ and t^* under a t-density with $n - 1$ degrees of freedom is $P/100$, where n is the sample size.

Calculation

- For a 99% confidence interval, we find $z^* = 2.58$.
- We need n so that
$$2.58 \times \frac{0.15}{\sqrt{n}} < 0.03$$
- Work on the board to show $n > \left(\frac{(2.58)(0.15)}{0.03} \right)^2 \doteq 167$.

General Formula

Sample size for a single mean

$$n > \left(\frac{(z^*)(\sigma)}{M} \right)^2$$

where M is the desired margin of error.

Rejection region

Example

- Graph the power for a sample size of $n = 25$ for $\alpha = 0.05$ for $H_0 : \mu = 3.35$ versus $H_A : \mu \neq 3.35$.

- Note that the p-value is less than 0.05 approximately when

$$\left| \frac{\bar{X} - 3.35}{0.15/\sqrt{25}} \right| > 1.96$$

- The rejection region is

$$\bar{X} < 3.35 - 1.96 \left(\frac{0.15}{\sqrt{25}} \right) \doteq 3.291$$

or

$$\bar{X} > 3.35 + 1.96 \left(\frac{0.15}{\sqrt{25}} \right) \doteq 3.409$$

Power when $\mu = 3.3$

- To find the power when $\mu = 3.3$, we need to find the area under the normal density centered at $\mu = 3.3$ in the rejection region.
- Here is an R calculation for the power.

```
> se = 0.15/sqrt(25)
> a = 3.35 - 1.96 * se
> b = 3.35 + 1.96 * se
> c(a, b)

[1] 3.2912 3.4088

> power = pnorm(a, 3.3, se) + (1 - pnorm(b, 3.3, se))
> power

[1] 0.3847772
```
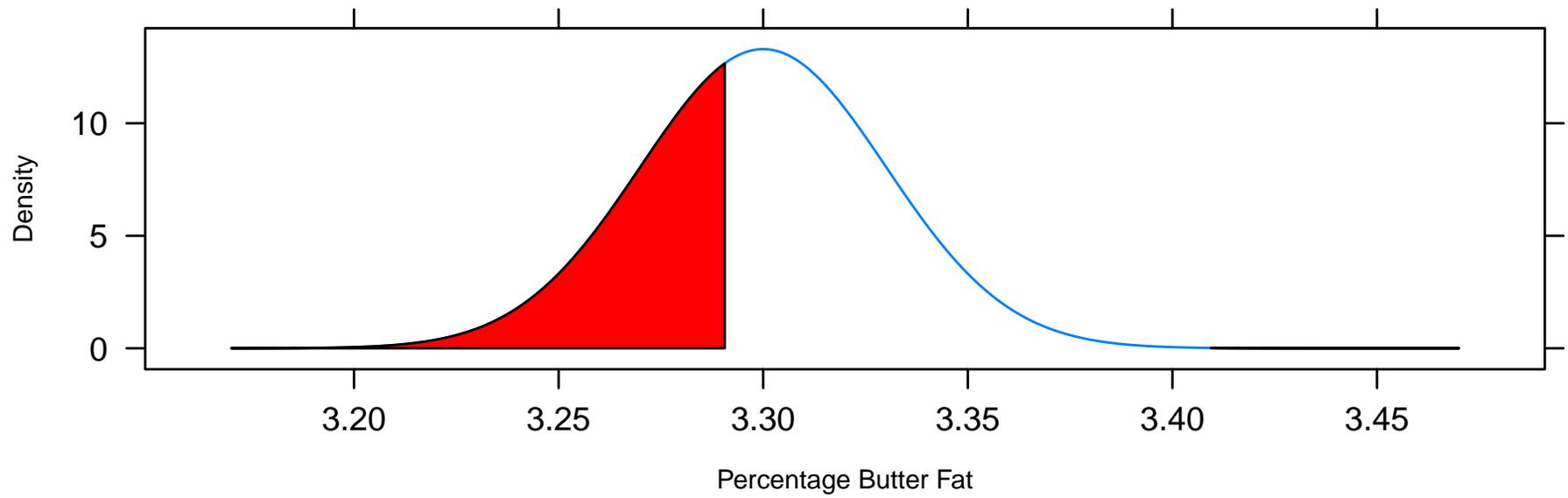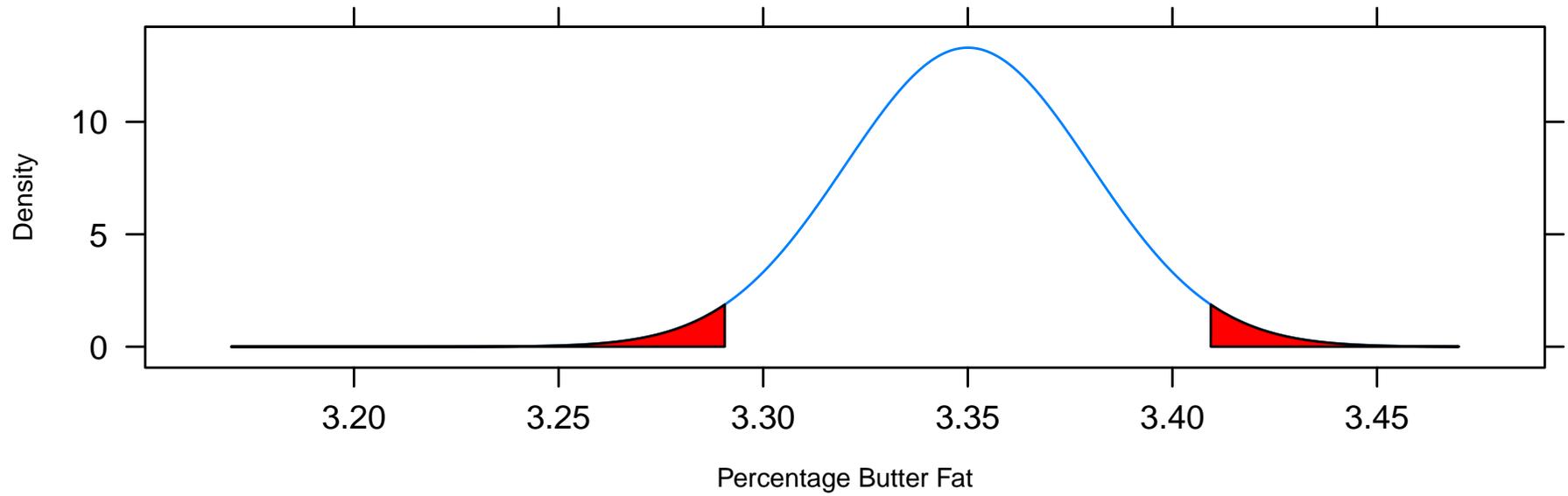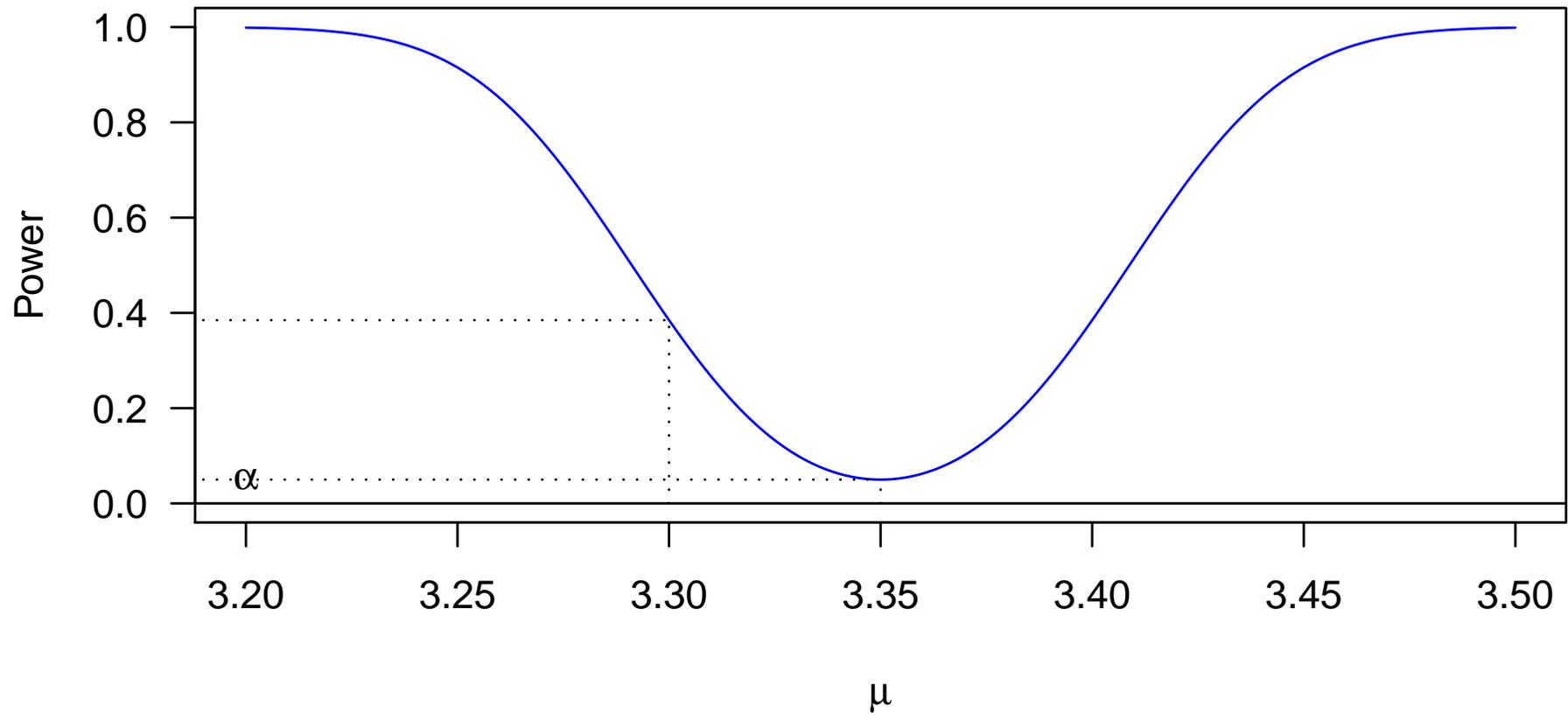
Graph of Power Calculations

- The previous slide shows a numerical calculation for the power at a single alternative hypothesis, $\mu = 3.3$.

- The next slide illustrates both the rejection region (top graph) and the power at $\mu = 3.3$ (bottom graph).

- The following slide shows a graph of the power for many alternative hypotheses.

Power: $n = 25$, $H_0 : \mu = 3.35$

Sample Size Calculation

Problem

- Long-run average percent butterfat in milk at a farm is 3.35 and the standard deviation is 0.15, with measurements taken by the load.

- How large should a sample size be to have an 80% chance of detecting a change to $\mu = 3.30$ at a significance level of $\alpha = 0.05$ with a two-sided test?

Solution

- The left part of the rejection region will have form

$$\bar{Y} < a = 3.35 - 1.96\left(\frac{0.15}{\sqrt{n}}\right)$$

  as $z = -1.96$ cuts off the bottom $\alpha/2 = 0.025$ of the area.
- The power when $\mu = 3.30$ is essentially just the area to the left of a under the normal density centered at 3.30 as the area in the right part of the rejection region is essentially 0.
- For the power to be 0.80, a must be the 0.80 quantile of the Normal$(3.30, 0.15/\sqrt{n})$ density.
- As the 0.80 quantile of a standard normal curve is $z = 0.84$, this means that

$$a = 3.30 + 0.84\left(\frac{0.15}{\sqrt{n}}\right)$$

- Set these two expressions for a equal to one another and solve for n.
- (Finish on the chalk board to verify $n \geq \left(\frac{(0.15)(1.96+0.84}{3.35-3.30}\right)^2 \doteq 71$.)

What you should know

You should know:

- what the definitions of power and significance level are;
- how to find sample sizes for desired sizes of confidence intervals for both proportions and means;
- how to find sample sizes with desired power for specific alternative hypotheses for both proportions and means;
- how to examine and interpret a power curve;
- how power curves change as n changes.