A Statistics Pot Luck

Charlie Gibbons
ARE 212

Spring 2011

Outline

Outline

Empirical CDFs

As we have seen in class, the empirical cumulative distribution function (ECDF) is important.

Let's consider a plot of it for the growth rate (roughly, log income) using the `tbrate` data set.

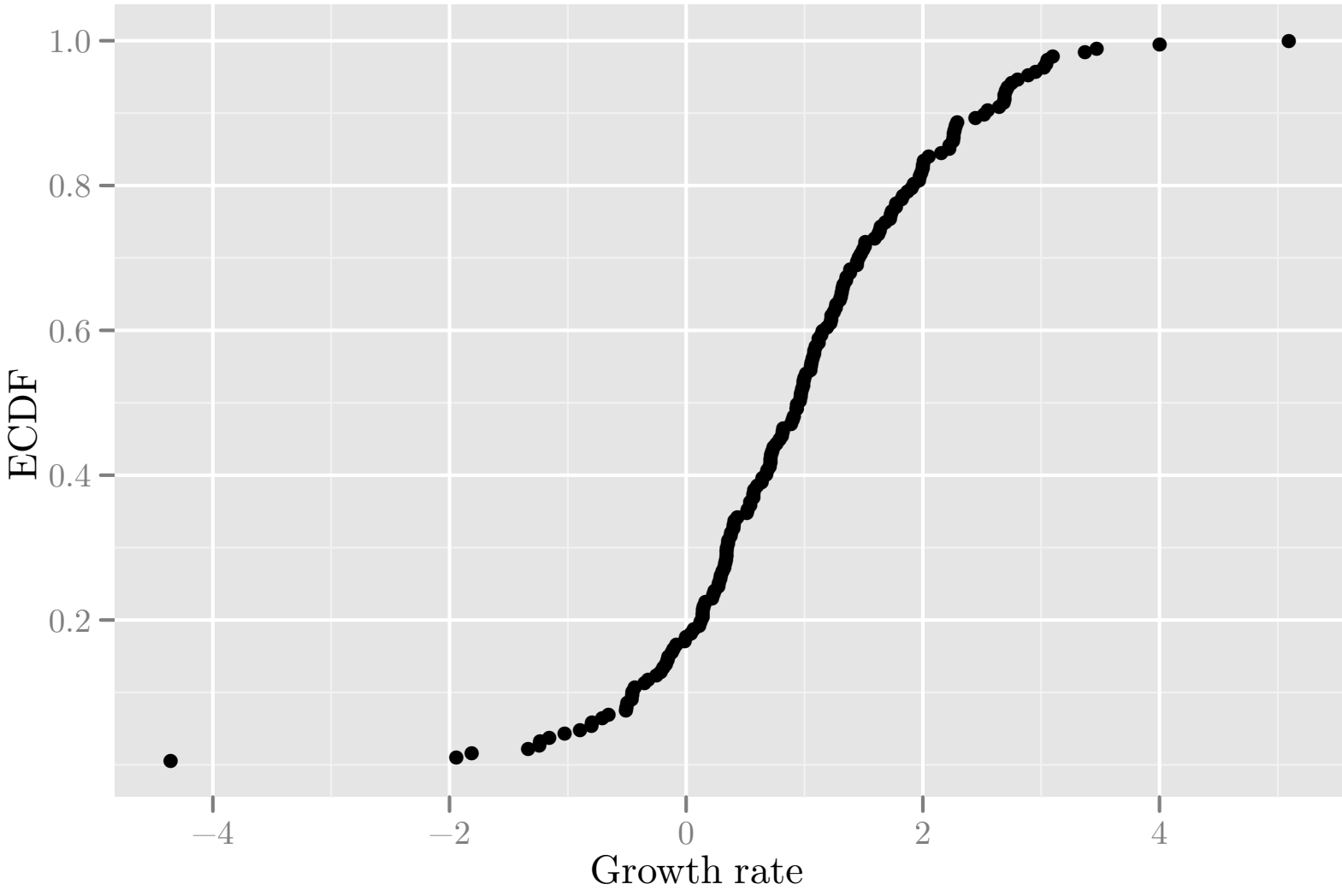

Figure: ECDF of economic growth rate (percent)

Comparison to a standard normal

This doesn't look like a standard normal CDF; the mean is different (*i.e.*, not 0) for one thing.

But how different is it?

A *QQ-plot* places a dot in the xy plane representing the percentile for each distribution. An example is the best explanation.

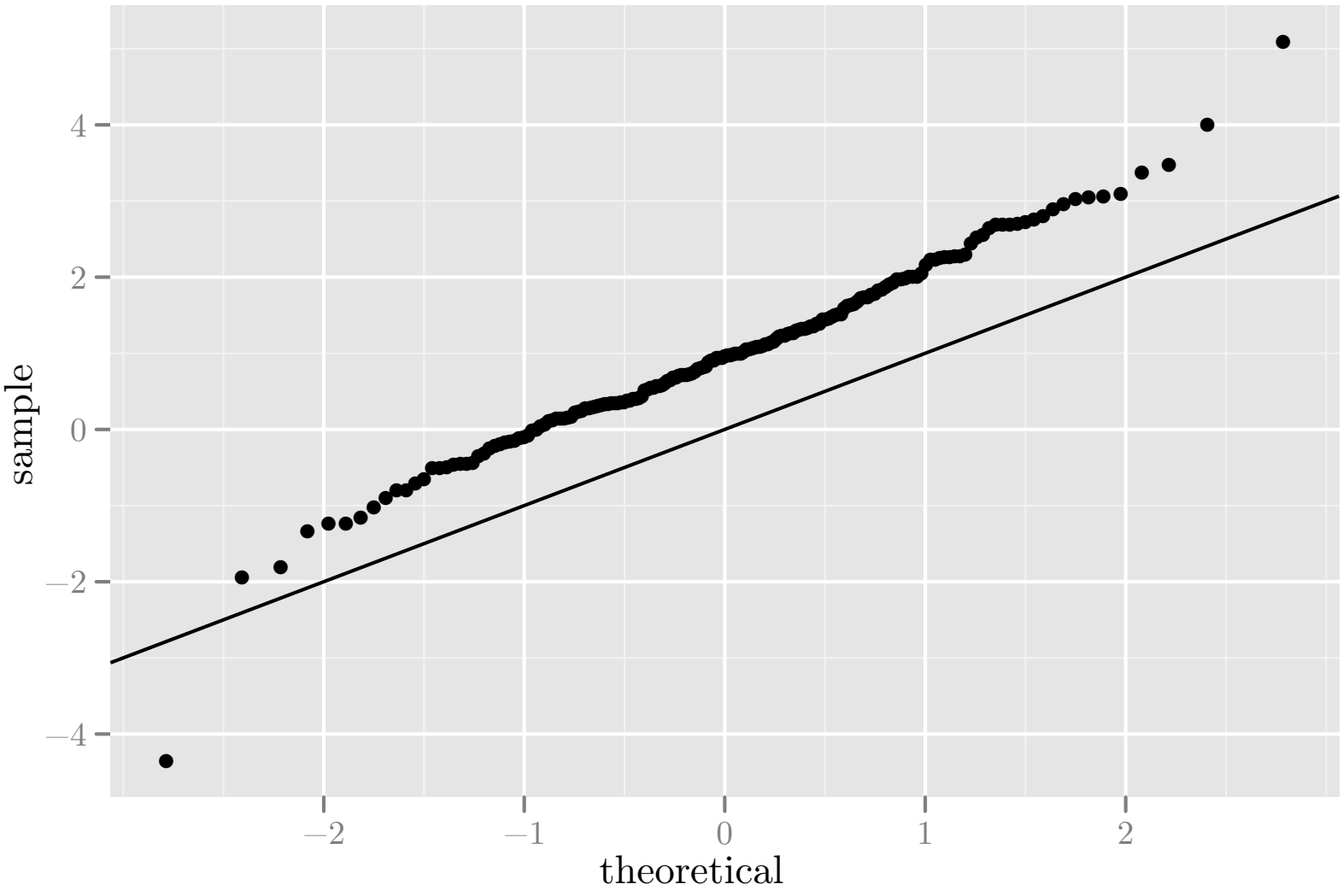

Figure: ECDF of growth rate versus a standard normal

Summary of QQ-plot

This picture reveals a

- Higher mean and a
- Larger variance/fatter tails.

Let's see some clearer examples.

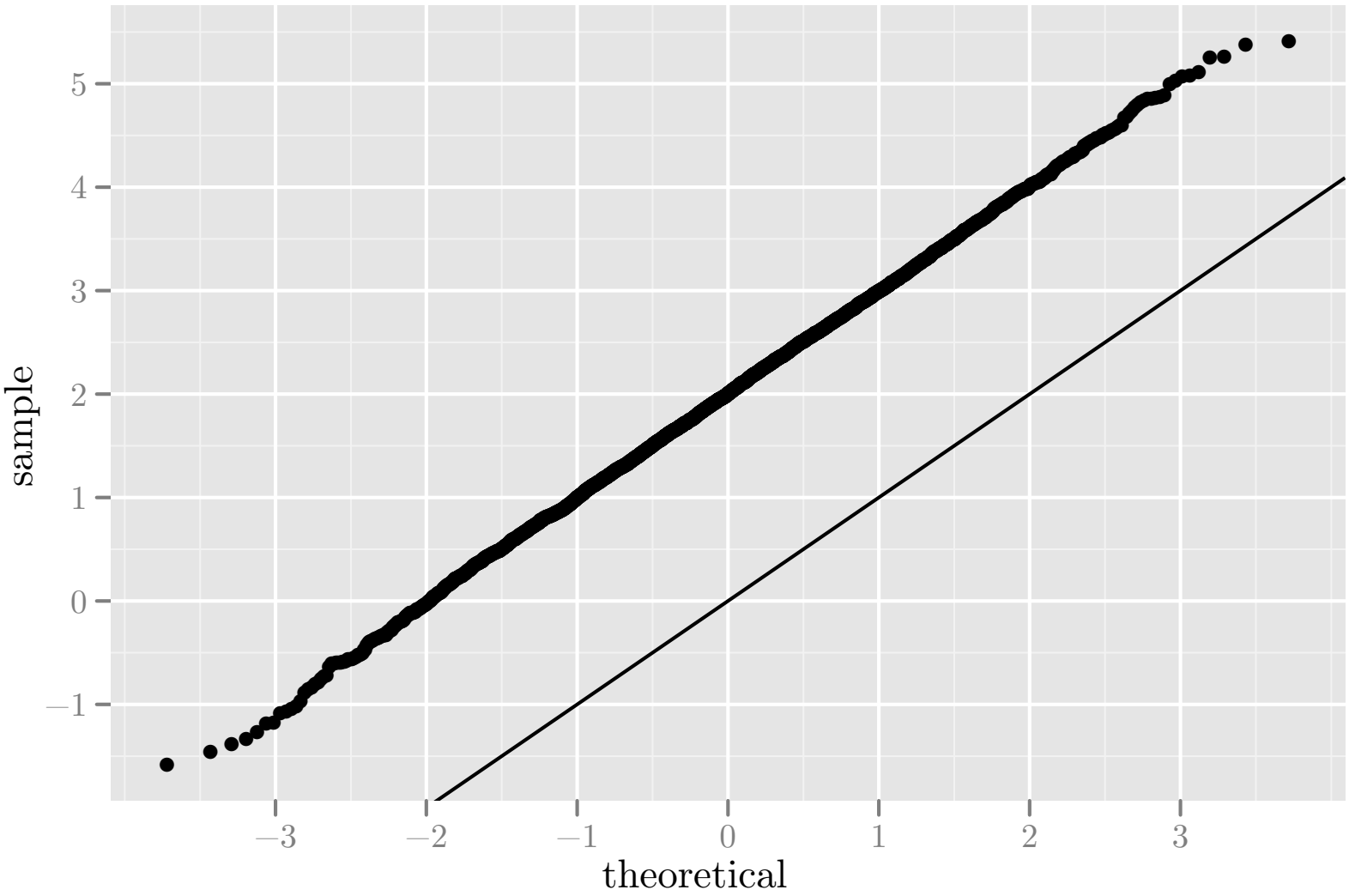

Figure: Normal distributions with the same variance, different means (0 *v.* 2)

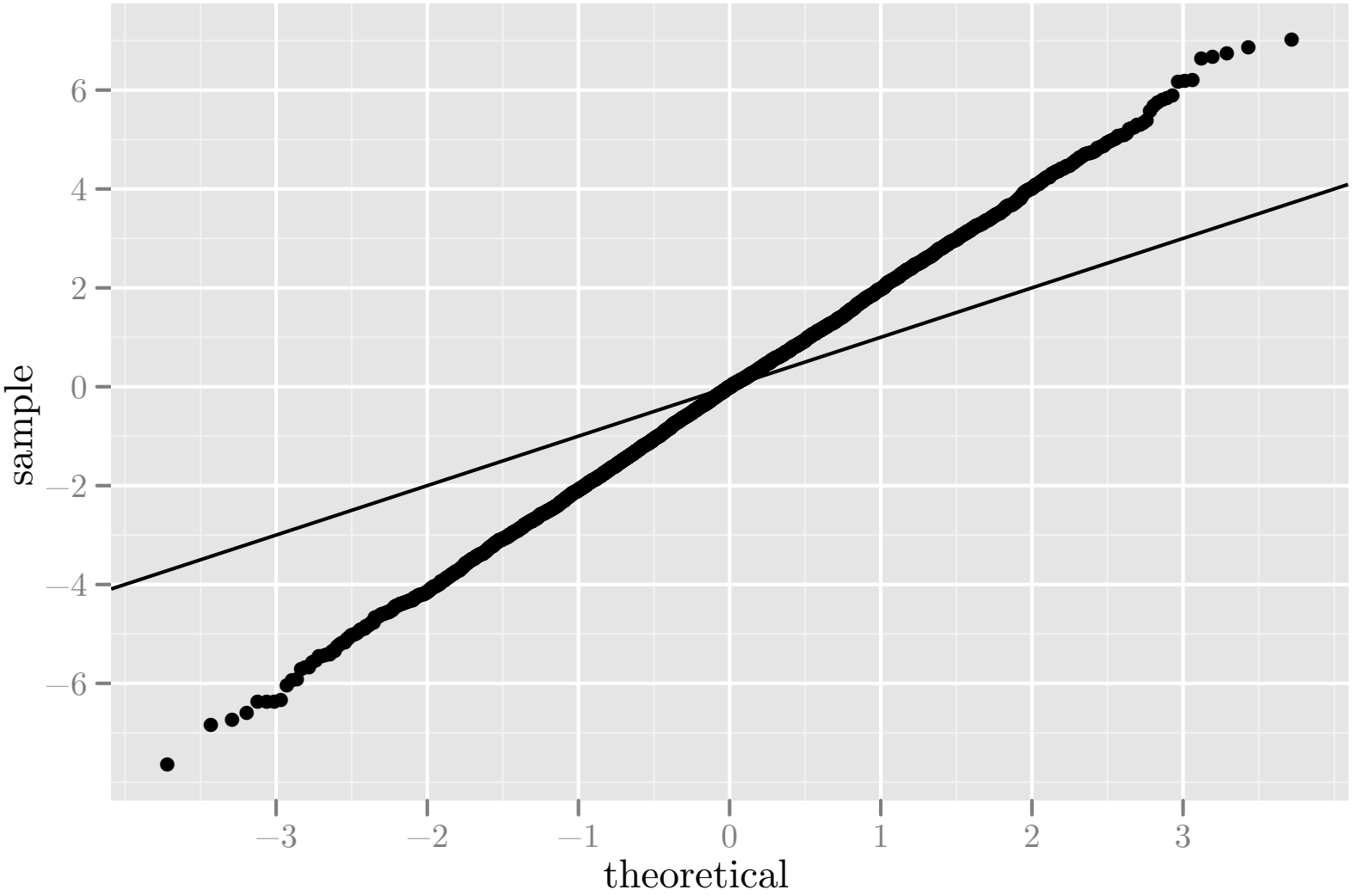

Figure: Normal distributions with the same mean, different variances (1 *v.* 2)

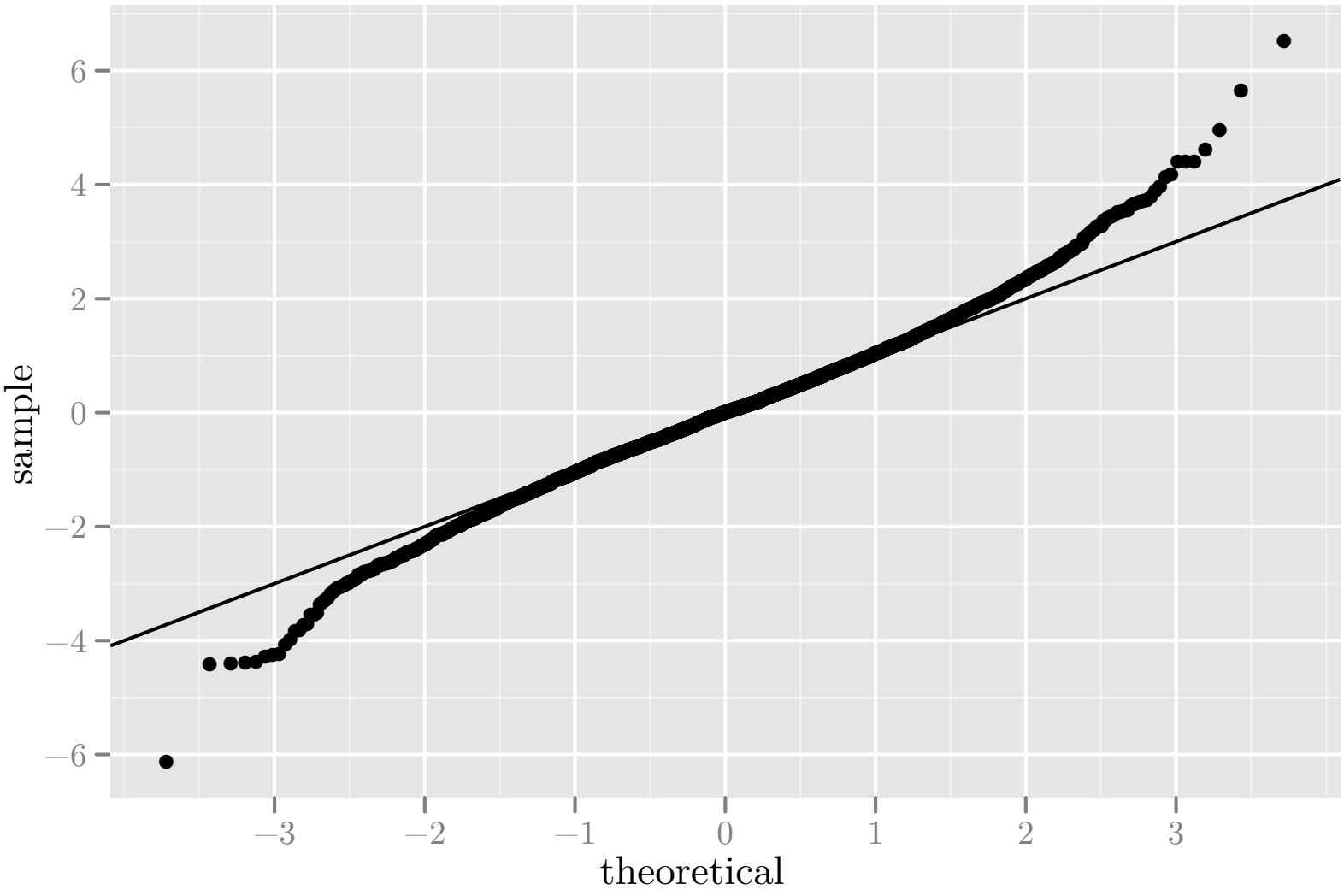

Figure: A distribution with fatter tails, but the same mean and variance as a standard normal (t with 10 D.F.)

KS test

These figures give an optical test of difference between distributions, but what statistic could we use to summarize the difference?

We can use the maximum absolute difference between the distributions, which is the maximum absolute (vertical) distance from a point to the 45-degree line.

This statistic is used in the non-parametric *Kolmogorov-Smirnoff (KS) test* for differences in distribution.

Outline

Minimizing mean squared error

To find $\hat{\beta}$ using OLS, we find the b that minimizes the mean squared error:

$$\min_b \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i' b)^2.$$

We have seen that outliers can exert a great deal of leverage and alter the regression results greatly.

This is because the mean is very sensitive to large quantities.

Minimizing least median of squares

Instead of minimizing the MSE, we can minimize the *median squared error*:

$$\min_b \text{median}\left\{(y_i - x_i'b)^2\right\}.$$

The median is not sensitive to extreme values.

Breakdown points

The *breakdown point* of an estimator is the number of observations that you need to change in order to get an arbitrary result.

Suppose that we want to change observations in x_i such that the mean of x is equal to some number K. Then,

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

$$\implies x_1 = K - \frac{1}{n}\sum_{i=2}^{n} x_i;$$

we can change just the value of x_1 to that given above to get a mean of x equal to K. The breakdown point of the mean is 1.

Breakdown point of the median

On the other hand, if we want to change the median of an observation *to any possible value*, then we would have to move half the data.

For example, if $x_i < K \, \forall \, i$, then we'd have to move half the data to be above K. The breakdown point of the median is $n/2$.

This means that, as long as at least half of your data are "correct" and aren't outliers, then the median is robust to the outliers.

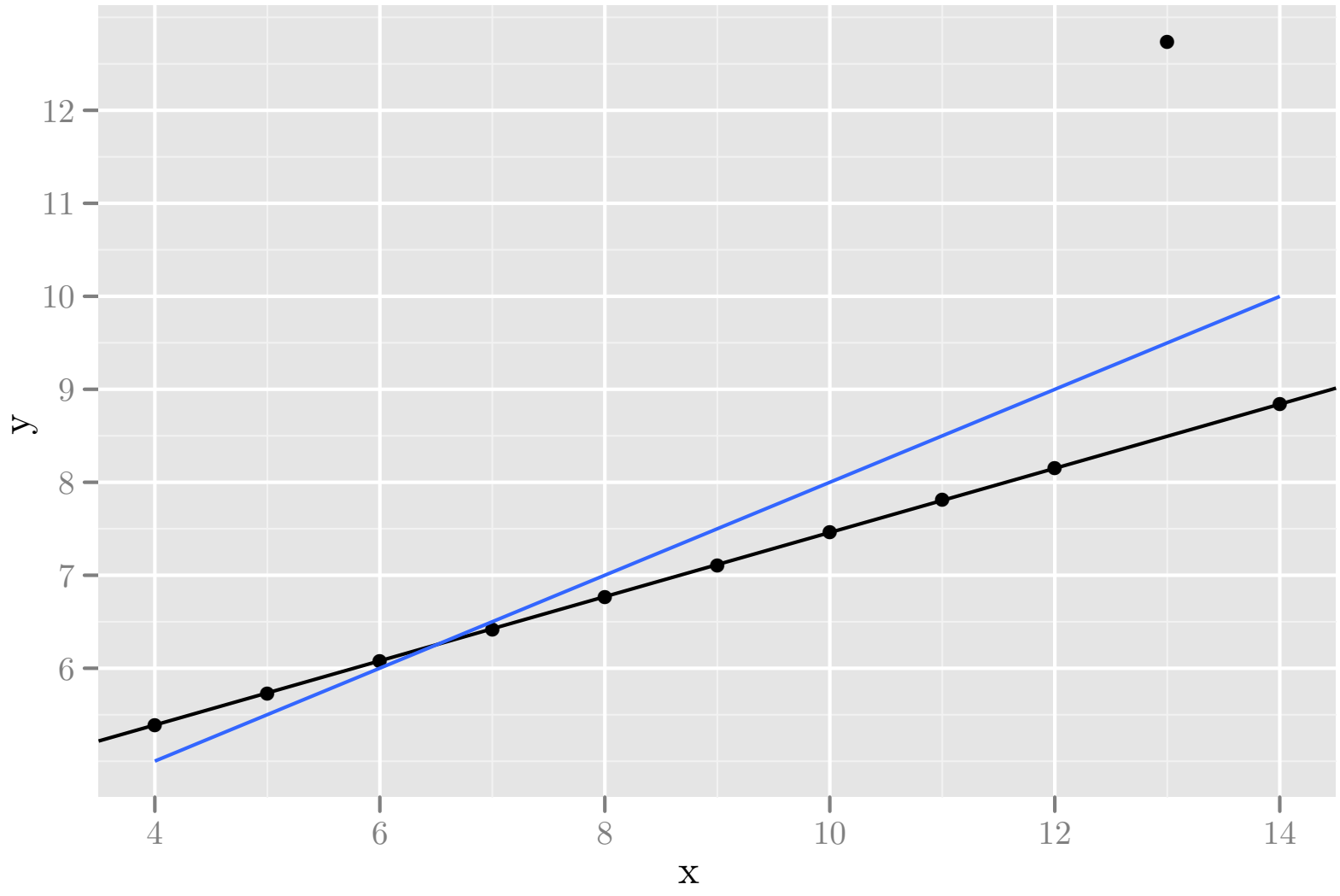

Figure: OLS (blue) and LMS (black) on data with an outlier

Downsides to robust regression

Why don't we always use robust regression?

- No closed-form results for $\hat{\beta}$ or standard errors, meaning that
- It can be computationally-intensive to compute (since the median is not a differentiable function), and
- The standard errors are bigger than for OLS.

To do LMS, we can combine numerical optimization and *bootstrapping* procedures.

Outline

Voting by rich people

Question: Are richer people more likely to vote Republican?

We're going to use median state income and the state percent vote share of then-Senator Obama from 2008 to answer this question.

Regression results

We regress the share of the 2008 presidential vote going to Senator Obama on the median income of the state:

Slope coefficient:	0.61
Intercept:	20.08

Income is in thousands of dollars; vote share is in percentage points

Interpreting the results

This result goes against our intuition. That's because *it's wrong* (see, *e.g.*, , Gelman et al. 2007).

What is the correct way to interpret these results?

Ecological fallacy

These results suggest that wealthier *states* are more likely to vote for a Democrat; it is incorrect to say that wealthier *people* are more likely to vote for the Democrat.

We cannot use inferences based upon grouped data to infer what individuals do; this is the *ecological fallacy*. Such analyses are subject to *aggregation bias*.

Consistency assumption

We can make an assumption so that individual behavior can be inferred from group behavior:

Consistency assumption

If the voting choice of an individual does not vary systematically with the median income in the state (or by state generally), then we can infer from group to individual behavior.

In other words, someone that earns $X in a rich state has to be just as likely to vote for a Democrat as someone who earns $X in a poor state.

Marginal data

Let's change the problem: Define high-income people as having income above the median income in New Hampshire ($65,000). Then, the ecological inference problem for that state is:

	Obama	McCain	
High income	???	???	50%
Low income	???	???	50%
	55%	45%	

We want to fill in those question marks.

Probabilities

We can translate this table into math:

$$\begin{aligned}\Pr(\text{Obama}) &= \Pr(\text{Obama} \mid \text{High income}) \Pr(\text{High income}) \\ &\quad + \Pr(\text{Obama} \mid \text{Low income}) \Pr(\text{Low income}) \\ &= 0.5 \Pr(\text{Obama} \mid \text{High income}) \\ &\quad + 0.5 \Pr(\text{Obama} \mid \text{Low income}) \\ \implies 0.5p + 0.5q &= 0.55,\end{aligned}$$

where $p = \Pr(\text{Obama} \mid \text{High income})$ and $q = \Pr(\text{Obama} \mid \text{Low income})$ and we implicitly condition on being in NH. We have one equation and two unknowns.

Calculating bounds

We rearrange this equation to get

$$p = 1.1 - q.$$

Since $0 \leq q \leq 1$, $0.1 \leq p \leq 1.1$, but, of course, $p \leq 1$ as well. Hence, we can say that the probability that a high income voter in NH cast a ballot for Senator Obama is between 10% and 100%.

The bound here is rather uninformative; the width of the bound is equal to the ratio of being in the high income group to the probability of being in the low income group. A length of 1, which we have here, is the widest possible. We get much better inferences when we have overwhelming majorities.

Point identification

Suppose that the consistency assumption holds. Keeping the same definition of high-income, if we go to another state, we get a second equation *based upon the same two unknowns* as the one that we already have and we can solve for p and q uniquely.

In fact, we can go to all 50 states, calculate the intervals for p and they should contain a unique intersection (this is called *overidentification*).

If there is no intersection, then we can refute the consistency assumption.

The one problem: We've assumed that we measured all our marginal quantities without error.

For more, see Freedman (1999) and Manski (2007).

Outline

Gender discrimination

Question: Does UC Berkeley discriminate against women?

Berkeley was sued for gender discrimination in graduate admissions in the 1970s.

We're going to use graduate admissions data from 1973 for Berkeley's six largest departments to asses this claim.

Discrimination!

Here is a summary of the data:

```
  Gender Admitted Rejected
1 Female      557     1278
2   Male     1198     1493
```

The acceptance rate for men is 44.5%; for women, it is 30.4%. The odds ratio is 1.83—men are 83% more likely to be admitted than women.

Is there something innocuous that could explain this result?

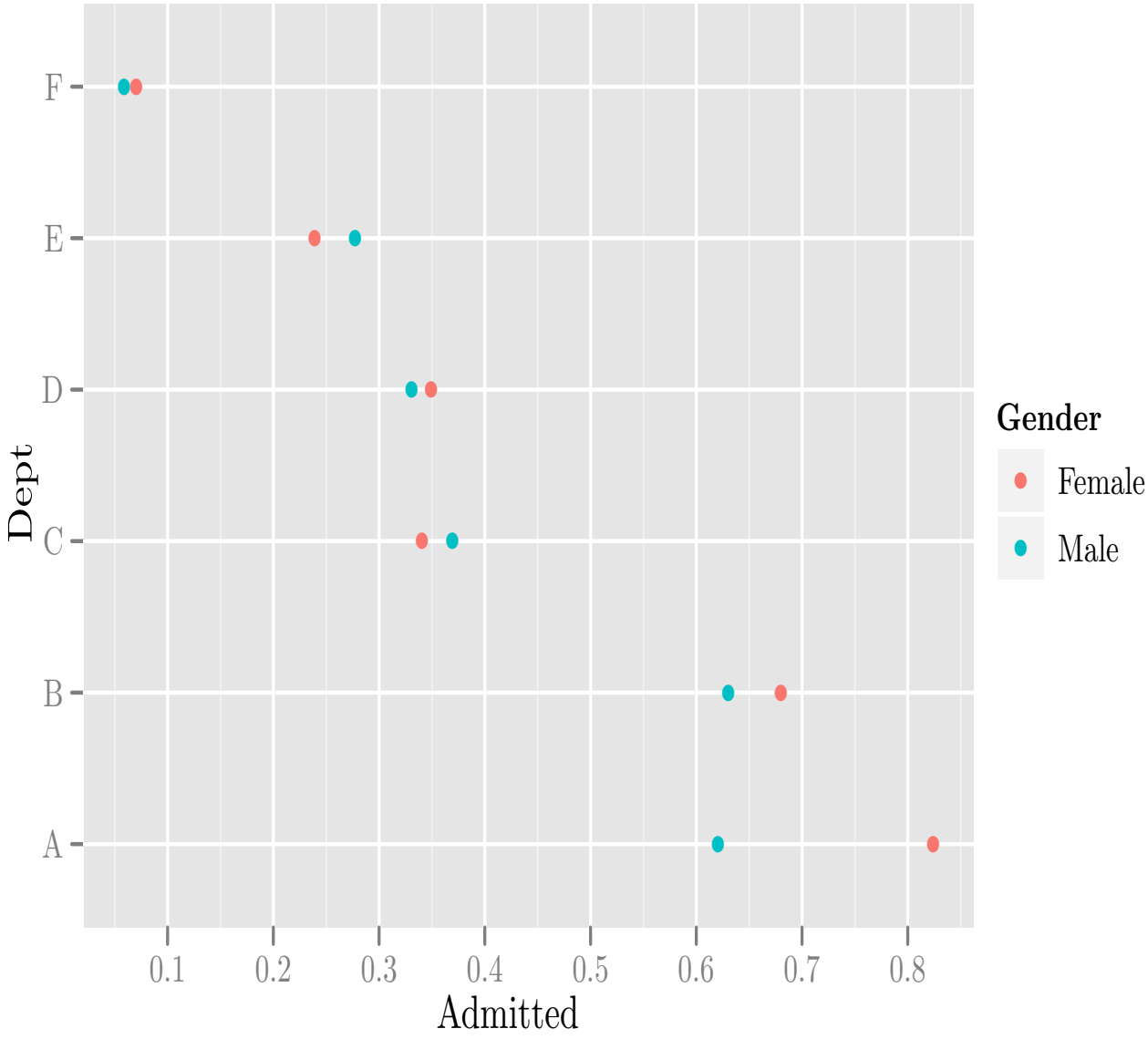

Figure: Admissions rates by department for men and women

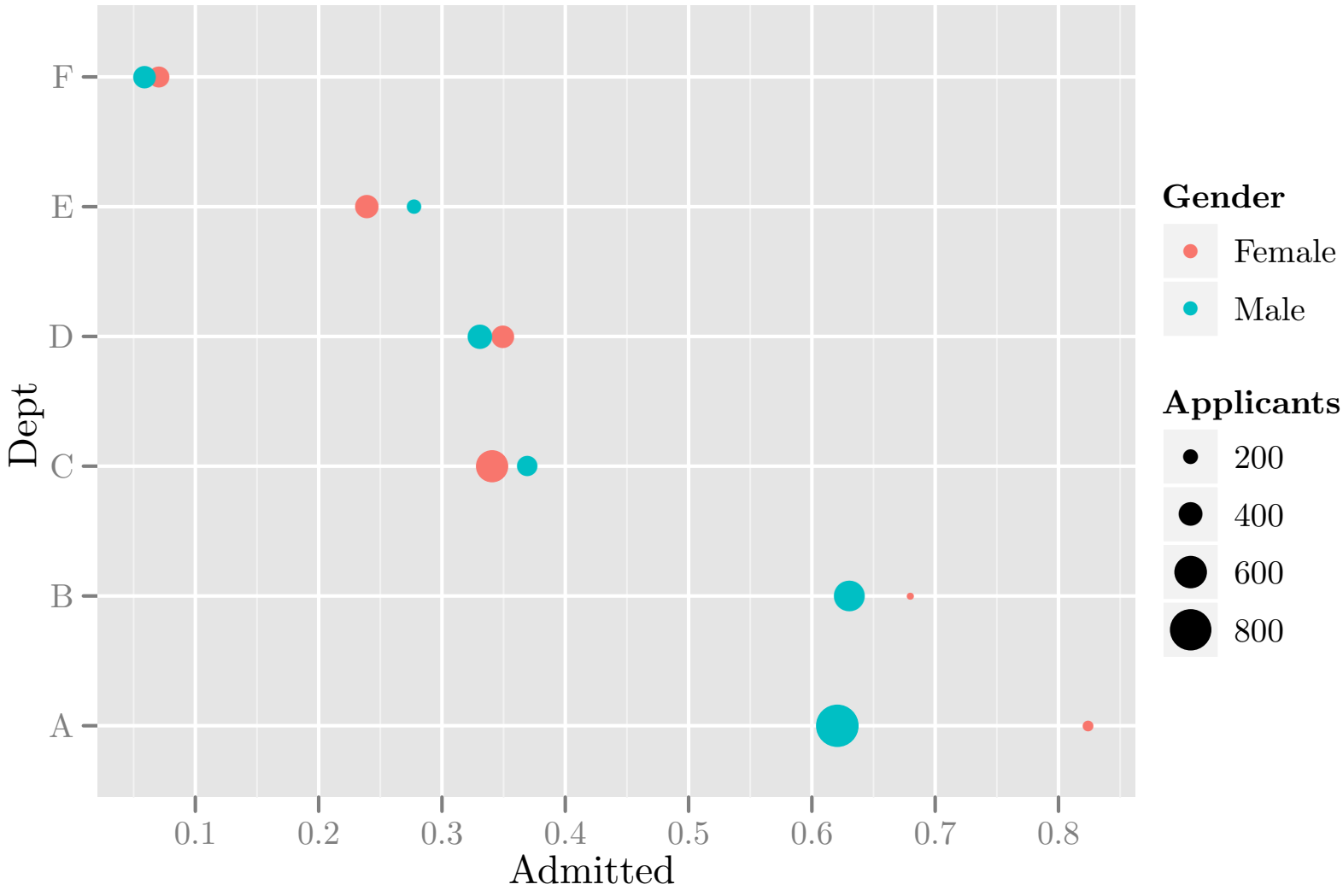

Figure: Admissions rates by department for men and women

Simpson's paradox

Women tend to have higher admissions rates, yet their overall admissions rate is lower than it is for men.

The issue is that women disproportionally apply to competitive departments and men apply to less competitive departments.

This is known as *Simpson's paradox*.